INTERSTELLAR MISSION COMMUNICATIONS
LOW BACKGROUND REGIME *

PHILIP LUBIN,1 DAVID G MESSERSCHMITT,2 AND IAN MORRISON3

1 University of California at Santa Barbara, Department of Physics, USA
2 University of California at Berkeley, Department of Electrical Engineering and Computer Sciences, USA
3 Swinburne University of Technology, Centre for Astrophysics and Supercomputing, Australia

(Received January 20, 2018)
Submitted to ApJ

ABSTRACT
Current attention on interstellar probes for near-term exploration of nearby star systems is focused on low-mass probes that can be accelerated to relativistic speed using propulsion from a ground-based DE beam. We consider the design of an optical communication downlink for the return of scientific data from such a probe at the distance of Proxima Centauri. The conditions under which background radiation can be neglected are quantified, and the design operates within that regime. Direct-detection is preferable to heterodyne, and in that context the transmitter should attain high peak-to-average transmitted power ratios. Based on available electric power sources, the downlink is expected to operate for years or even decades following target encounter, combined with low data rates. There are several areas in which technology innovations are needed, most of them related to Earth-based large-area aperture receiver design with direct detection. A major issue is the choice of multiplexing approach to support multiple probe downlinks and related challenges. Due to the interaction of trajectory parallax effects with field-of-view, we conclude that aperture synthesis with controlled optical beam forming may be required to reject radiation from the target star. Short visible wavelengths for laser communications are also highly advantageous in reducing that radiation. Highly selective optical bandpass filtering is needed to reject unnecessary background radiation, and a short-term data uplink is required to configure the transmit wavelength for variations in probe speed. Fundamental limits on the photon efficiency are compared to a concrete modulation/coding design in the presence of weather-based outages.

Keywords: interstellar, communications, probes

Corresponding author: David G Messerschmitt
messer@eecs.berkeley.edu
* Copyright c⃝2018.
1. INTRODUCTION

Both interest and concrete efforts toward interstellar space probes for scientific investigation of other solar systems are growing. To reach even the nearest stars and recover useful scientific data within a human lifetime requires relativistic speeds. With our current state of technology, low-mass probes propelled by directed energy (DE) is the most promising approach.

Current thinking centers around a probe traveling at 10% to 20% of the speed of light and thus requiring 20 to 40 years to reach the vicinity of nearby stellar system such as Proxima Centauri as in the NASA Starlight and Breakthrough Starshot programs (Lubin 2016). The communication of scientific data back to Earth is the subject of this paper. We focus on the downlink (probe to Earth) design with the expectation that any uplink (Earth to probe) communication activity is likely to be short-lived and and benefit from relative proximity to Earth (for functions like wavelength-configuration and course-correction).

There is a tradeoff between probe mass, speed, data rate, and other parameters. Here we focus on the low-mass “wafer scale” probe version, which is the most challenging of the design cases because transmit power and transmit aperture size are quite restricted. Tradeoffs between higher mass (with commensurately lower speed and longer transit time) and higher data rate (and thus greater scientific return) are also discussed.

Relying heavily on a reusable fixed propulsion and communication infrastructure confined to the Earth’s surface, a swarm of multiple probes could be launched. Following flybys of the target star and its exoplanets, scientific data representing images, magnetic fields, spectroscopy, etc. would be returned to Earth following the target encounter.

In some ways this challenge mirrors the extensively studied requirements for interplanetary communication within our solar system (Hemmati 2006, 2009). The biggest differences include the severe limitations on the probe’s available electrical power, processing, and transmit optics, as well as the much greater propagation loss, different sources of interfering background radiation, ground-based reception with many issues related to atmospheric turbulence and scattering and weather, and the expectation of multiple probe downlinks operating concurrently. If crippling limitations to data volume are to be avoided, careful attention must be paid to rendering the communication link as energy efficient as possible, to aggressive compression of the data, and to the use of receive optics with much greater aperture size than has been achieved previously. If optics is shared between DE and communication, this introduces additional design challenges.

We develop a framework for the design of an optical downlink with its numerous building blocks, and develop a preliminary design approach within that framework. The emphasis is on a holistic treatment of the design with coordination and interoperation of the constituent elements. The primary goal is to identify the “points of pain” where feasibility is in doubt and invention and technology advances are needed. Rather than make definitive design choices, we stress and quantify the tradeoffs that will guide those choices. This helps guide and focus future research activities in support of such a mission.

In this paper we simplify the downlink design by assuming that background radiation can be neglected. Rendering this assumption valid through choice of system parameters such as electrical power, optics, bandwidth, and data rate should be a design goal, since background radiation would otherwise be deleterious to performance metrics. One outcome is therefore to quantify the range of these parameters that fit this assumption. The background-free assumption desirably yields closed-form analytical results that permit simple and intuitive design exploration, and is thus valuable in arriving at a baseline design.

It is also important to understand the ramifications of allowing material background radiation, which will permit greater design freedom in the choice of major system components and performance objectives. Dealing with background radiation forces a more cumbersome numerical approach, and will be pursued in a follow-on paper.

Today most interest in interstellar exploration by space probes resides in the astronomy and astrophysics communities, but some essential knowledge and experience lies with the communications sciences. With the goal of informing all these communities and guiding them around this challenge, we emphasize intuition and explanation.

1.1. Previous work

Space communication applications have stimulated extensive research and engineering into long-distance free space optical communication. Results most relevant to the present challenge come from JPL’s interplanetary network (IPN) (Dolinar et al. 2012).

Interstellar communication differs from terrestrial communication in its emphasis on energy efficiency rather than spectral efficiency. This distinction is more evident at radio wavelengths, where energy-efficient interstellar communication has been studied in the context of METI/SETI (Messerschmitt 2013, 2015). However, sophisticated error-control coding is presumably not feasible for initial contact with extraterrestrial civilizations due to a lack of coordination between transmitter and receiver, but always plays an essential role in communication with space probes.

The germane sources of optical-wavelength background radiation for interstellar communication are a current area of research (Lubin et al. 2018). Our previous background calculations have been independently quantified in (Hippke 2017).
1.2. Comparison to Solar System missions

To get a preliminary idea of the viability of such a mission, it is informative to compare an interstellar mission with the New Horizons spacecraft, which performed a recent exploration of the outer Solar System. Some parameters are listed in Table 1, with only rough figures for a hypothetical interstellar mission listed (one goal of this paper is to fill in some more definite values).

The biggest obstacle to overcome is the approximately 6800 AU greater propagation distance compared to Pluto, which results in a 4.6·10^6 greater power loss. A low-mass interstellar probe has two other obstacles, which are a smaller transmit aperture and smaller transmit power. Several other parameters can be adjusted to compensate for distance:

- Shorter wavelength (optical rather than radio) reduces loss for a given aperture size.
- A significantly larger receive aperture reduces loss.
- A lower data rate reduces the receive power requirement, and a longer transmission time can expand total data volume returned.
- A swarm of probes (1000 or more) multiplies both the scope of observations and the scientific data return for a given data rate.

Our emphasis is on the scaling laws that govern design tradeoffs. A convenient “stake in the ground” for numerical results is a nominal data rate of \( R = 1 \) bps, but we also quantify the scaling laws which permit an increase or decrease in this rate. Receive aperture diameter is particularly significant, as an assumed range of 1-10 km diameter corresponds to a 100x range in \( R \).

1.3. Quantum limits

This paper follows the conventional approach of photon-counting detectors, which are well-supported by current technology and many concrete system implementations. Performance can be improved in theory by generating and querying multiple quantum states concurrently. While the fundamental limit on the performance of such hypothetical sources and detectors has been obtained (Giovannetti et al. 2004; Dolinar et al. 2011), no concrete technology has yet been proposed or developed which improves on the classical techniques pursued here. Nevertheless it is worthwhile to quantify the quantum limits on communication for this application (Hippke 2017) as this indicates the feasible future improvement with technology advances.

2. DOWNLINK BUILDING BLOCKS

2.1. Scientific objective

Many types of scientific instrumentation may be feasible. The primary relevance of scientific objective to the downlink design is the total data volume \( V \) to be communicated and the reliability objective (as measured by a probability of error \( P_e \)). Early missions are likely to emphasize imaging, so this is the application emphasized in the design to follow.

2.1.1. Data volume

A reasonable assumption is 1-3 Mb of data to represent a compressed image. These figures will accommodate 10^6 pixels with 1-3 bits per pixel. The lower value may accommodate high-quality images considering the prevalence of a black background in many of the images. At a nominal \( R = 1 \) bps data rate, the transmission time for a 1 Mb image would be 11.6 days. Of course that transmission time is inversely proportional to data rate, and would be three times greater for a 3 Mb image.

Relative to solar system missions, the number of scientifically useful images will be reduced by the flyby at relativistic speed and expanded by the swarm of probes. For example 30 images per probe, or a totality of thousands of images, may yield a significant scientific return.

2.2. Probe trajectories

A flyby trajectory to a nearby star system such as Proxima Centauri is assumed, with a terrestrial receiver for the scientific data downlink. The mission parameters are listed in Table 2 and parameters particular to the downlink are listed in Table 3. Numerical examples to follow are based on the listed values.

A number of other mission scenarios discussed in (Lubin 2016) would make substantial modification to the assumptions in Table 3. For example, the laser “sail” (∼ 1 m diameter) might be used to increase the transmit aperture area.

2.2.1. A swarm of probes

The cost of a low-mass probe launch system is concentrated in the Earth-based DE and communications infrastructure. As a result the incremental cost of each probe launch is not as significant. A swarm of multiple

\[ V \text{ in bits} (b) \text{ rather than bytes} (B). \text{ For example,} \]

\[ 10^6 \text{ bits} (125 \text{ kilobytes}) \text{ would be written as} 1 \text{ Mb}. \]
probes can increase the scientific return by increasing the coverage of the target star system, or even multiple star systems. Reliability is also improved through redundancy.

2.2.2. Trajectory differences

There is a strong economic motivation to service all probes with a single shared receiver. The required field of view (FOV) of that receiver must take into account the differences in trajectories among those multiple probes. This is illustrated in Figure 1. The angle of each trajectory as seen from Earth varies seasonally due to the transverse movement of Earth, which is a parallax effect. In addition, each probe will have a finite-size target bullseye accounting for any imprecision in the launch vector. The position of that bullseye may differ among probes as influenced by scientific objectives (for example a desire to image different exoplanets).

2.3. Source and propagation

2.3.1. Detected power

\( \Lambda_A \) equals the average detected photons per second (rather than Watts). \( \Lambda_A \) depends on all the optical and photonic physical elements in the transmitter and receiver, as well as propagation in interstellar space and through the Earth’s atmosphere. Collectively these effects are called the physical layer.

Specifically, neglecting interstellar absorption and atmospheric absorption and turbulence, \( \Lambda_A \) is determined by the average transmitted power \( P_A \) (in Watts), the net attenuation factor \( \alpha \) due to optics and propagation, and the wavelength \( \lambda_0 \) through

\[
\Lambda_A = \frac{\lambda_0}{h c} \cdot \alpha \quad \text{where} \quad \alpha = \frac{\eta A_T A_R}{\lambda_0^2 d^2}.
\]

The factor \( \lambda_0/hc \) converts from Watts to photons per second. The freespace attenuation factor \( \alpha \) follows a standard result for transmit and receive aperture areas \( A_T \) and \( A_R \) and aperture area-product \( A_T A_R \). The efficiency factor \( 0 < \eta \leq 1 \) accounts for non-propagation losses, such as pointing error, attenuation in the receive optics, and optical detector efficiency.

| Description                                  | Value     |
|----------------------------------------------|-----------|
| \( D_t \) Distance of Earth to target star   | 4.24 ly   |
| \( D_d \) Distance from target star to termin- | Variable  |
| ation of downlink transmission               |
| \( D_b \) Variation in transverse distance at | 2 AU      |
| launch and during reception due to Earth     |
| motion                                       |
| \( D_e \) Radius of a target bullseye        | Variable  |
| \( v_0 \) Speed of probe                    | 0.2c      |

Figure 1. A simplified 2-D model of the transverse variation in probe trajectories across multiple probes due to (a) the finite baseline for both launch and reception (the Earth’s orbital motion) and (b) a finite target bullseye due to scientific objectives and launch aiming inaccuracy. The two probes #1 and #2 at the most extreme angles are shown, defining the required FOV. The probe motion is assumed to be rectilinear (neglecting gravitational influences) and the transverse direction is exaggerated in scale.

2.3.2. Wavelength

We assume a terrestrial receiver (as opposed to Earth orbit or on the Moon) for cost and operational reasons. The atmosphere is suitably transparent at microwave and visible-infrared optical wavelengths. The \( d^2 \) term in (1) is about \( 4.6 \cdot 10^7 \) larger for the interstellar mission. This can be overcome in part by choosing a shorter wavelength \( \lambda_0 \). As a starting point we can choose an infrared wavelength \( \lambda_0 = 1 \mu m \). The impact of a change from 3 cm (New Horizon mission) to 1 micron (our current baseline) depends on which measure of power we use:

- In Watts (more appropriate for a heterodyne radio receiver) the advantage is \( \propto \lambda_0^{-2} \), or \( 9 \cdot 10^8 \).
- In photons per second (more relevant to a direct detection optical receiver) the advantage is \( \propto \lambda_0^{-1} \), or \( 3 \cdot 10^4 \).

Shorter wavelength increases the precision required of optics elements, and in addition greater impairments due to atmospheric effects (absorption, turbulence, scattering, and weather events).

An additional distinction between microwave and optical is the nature and size of background radiation. Shorter wavelengths yield a significant reduction in unwanted background from a target star (Section 5). Ideally going to a space based system, at least for the receive array would be extremely advantageous (except
Table 3. Nominal transmission design parameters

| Description                  | Value                  |
|------------------------------|------------------------|
| $m_0$                        | Variable               |
| $P_A$                        | Average TX power (W)   | 10 mW                 |
| $\Lambda_A$                  | Average rate of detected photons (s$^{-1}$) | Variable |
| $P_P$                        | Peak TX power (W)      | 10.24 W               |
| $\Lambda_P$                  | Peak rate of detected photons (s$^{-1}$) | Variable |
| $\Lambda_B$                  | Average rate of detected background (s$^{-1}$) | Variable |
| $A_T$                        | Transmit aperture area (m$^2$) | 100 cm$^2$            |
| $A_R$                        | Receive aperture area (m$^2$) | Variable |
| $\lambda_0$                  | Wavelength (m)         | 1.0 $\mu$m            |
| $\eta$                       | Detection efficiency   | Variable               |
| $B$                          | Measurement bandwidth (Hz) | Variable |
| $W$                          | Communication bandwidth (Hz) | Variable |
| PAR                          | Peak-to-average power ratio $(P_P/P_A)$ | 1024 |
| SBR                          | Signal-to-background average power ratio $(\Lambda_A/\Lambda_B)$ | Variable |
| BPP                          | Photon efficiency (bits per photon) | 10 |
| $\mathcal{R}$               | Reliable scientific data rate (bps) | Variable |
| $T_d$                        | Transmission duration (yr) | Variable |
| $T_l$                        | Latency, or time elapsed from launch to last received bit (yr) | Variable |
| $V$                          | Total data volume (bits) | Variable |

for cost) both for “no weather outages” but even more important is the possibility of going to mild UV laser comm (0.2–0.3 micron for example) to vastly reduce the host star background (Lubin et al. 2018). While UV laser comm is not currently technologically advanced it is an option to consider for the future.

2.3.3. Photon efficiency

Consider the decomposition of the data rate $\mathcal{R}$ (in bits per second, or bps) reliably communicated on the downlink into two factors,

$$\mathcal{R} = \text{BPP} \cdot \Lambda_A.$$  \hspace{1cm} (2)

The factor BPP is the photon efficiency in bits per photon. BPP depends on how data bits are mapped into a transmit intensity waveform, as well as the coordinated receiver processing of the detected photon counts and timing. Collectively these are part of the coding layer (Section 3). The coding layer is principally implemented through algorithms and processing, and achieving large BPP requires substantial processing overhead. Fortunately most of this burden falls on the terrestrial receiver.

In this paper the coding layer design and performance measurement is simplified by assuming a regime in which background radiation can be neglected. Section 5 considers the conditions under which this assumption may be valid. A future paper will consider the implications of significant background radiation.

2.3.4. Peak power

The peak power of transmission $P_P$ is determined by the laser source technology and transmit optics. The peak-to-average power ratio (PAR) is the ratio $\text{PAR} = P_P/P_A = \Lambda_P/\Lambda_A$. Photon efficiency BPP is fundamentally limited to (Section 3)

$$\text{BPP} < \log_2 \text{PAR} \text{ for } \text{PAR} > e.$$  \hspace{1cm} (3)

Neglecting background radiation, reliable data communication is feasible when (3) is satisfied, and high reliability cannot be achieved when (3) is violated.

Based on the current knowledge and technology, it should be feasible to approach this limit reasonably closely at the expense of considerable processing in the coding layer (Section 4). The transmit-side processing required to achieve high BPP can be performed in the vicinity of the target star at the same time as image compression (using additional photovoltaic power derived from the target star if needed).

Design of the physical layer to increase PAR is advantageous in terms of increasing the “headroom” available to increase BPP. For fixed data rate $\mathcal{R}$, a larger BPP can be traded for smaller $\Lambda_A$, and hence smaller aperture area-product $A_AR$. At the expense of mass, for example, multiple laser diodes may be “ganged” through a common optics. PAR = $2^{10} = 1024$ will be assumed in numerical examples, and consistent with this BPP ≤ 10 bits per photon.

2.3.5. Electrical power

We assume two power sources: a radioisotope thermoelectric generator (RTG) or forward edge ISM proton impact converter during the cruise phase and photovoltaic power from the target star during the encounter phase. At relativistic speeds, the proximity of probe to target star consistent with significant photovoltaic energy production will be short-lived (roughly an hour). Thus photovoltaic is available for scientific data collection and the processing requirements of layer data compression and coding layer. Transmission during the subsequent communication phase is assumed to be exclusively powered by an RTG or ISM proton impact converter.

2.4. Data volume and latency

In terms of scientific return, the parameters of greatest interest are data volume $V$ and data latency $T_l$ (defined in Table 3). That is, how much data can be returned, and how long do we have to wait for it to be returned in its entirety?

2.4.1. Impact of probe mass

A variety of probe versions with different masses can be launched with the same DE infrastructure. At the expense of a longer transit time, higher mass probes can yield higher data rates (Lubin 2016). The many uses of a
Data latency in yr
20
40
80
6
21
55
et al. 2018). The design parameters in Table 3 as well as numerical examples in this paper are estimated for a DE launcher, its deployment, as well as the benefits and issues of various mass missions are discussed in (Lubin et al. 2018). The Pluto scientific data return was 6.25 GB (Fountain et al. 2009).

The wafer-scale case considered here brings more serious design challenges such as large accelerations, high illumination flux on the reflector, dynamic control issues, small transmit aperture $A_T$, low average power $P_A$, and overall lower data rate $R = R_0$. Assuming that $P_A, P_T, A_T \propto \xi$ (the detailed scaling depends on the specifics of the spacecraft design) it follows from (1) and (2) that $R = \xi^2 R_0$. Thus the data rate increase is {1, 9, 100, 900, 10^4}, which is much larger than the transit time increase.

Increased mass tends to increase $T_i$ because of longer transit time. It also increases the potential $V$ because of larger $R$, which is due to larger power and transmit aperture area and lesser propagation loss due to more slowly receding distance (see Appendix A.1). $T_i$ is plotted vs $V$ in Figure 2. Lower-mass probes return scientific data sooner, but higher-mass probes can return a greater data volume.

The $T_i$ in Figure 2 is the time to receive all the scientific data, but most data will be available earlier. The elapsed time until first data arrives is

$$D_i(1/u + 1/c),$$

which is {25.5 yr, 32.2, 42, 53.9, 71.3} for the five cases. The total data volume is capped by the square-law increase in propagation loss during transmission. This volume cap as $T_i \rightarrow \infty$ scales as $\xi^{3/4}$, and is {546 Mb, 6.8 Gb, 106 Gb, 1.3 Tb, 19.8 Tb} for the five cases. For comparison, the volume/latency for two outer-planetary missions are listed in Table 4 and shown in Figure 2.

For each $V$ there is an optimum choice of $\xi$ that minimizes $T_i$. The resulting minimum $T_i$ vs $V$ is plotted as the dashed curve. For the range of parameters shown, the latency falls in the 30 to 80 year range and the data volume falls in the 100 Mb to Tb range, assuming an initial data rate of $R_0 = 1$ bps for the $\xi = 1$ wafer-scale probe.

### 2.4.2. Mixed masses

In practice it is likely that a variety of masses will be deployed. It makes sense to start this program with higher-mass probes, since they are less challenging to design, fabricate, and operate. However, once a range of masses becomes available, for each new target star it makes sense to start with low-mass “scouts”, and follow up with higher-mass probes with more instrumentation and greater data return. All probes can share a common DE launcher. Due to field-of-view challenges (Section 2.10) in the context of continuous operation, a separate receive aperture is likely required for each independent target star.

---

**Table 4.** Data latency/volume for planetary missions

| Target       | Mission       | Data rate | Latency | Data volume |
|--------------|---------------|-----------|---------|-------------|
| Neptune      | Voyager 2     | ~1 kbps   | 12 yr   | 18 Gb       |
| Pluto        | New Horizons  | ~1 kbps   | 10.5 yr | 50 Gb       |

---

---

2. Although probes with even higher mass can be brought to relativistic speed, one fundamental issue (required laser illumination time) favors space-based DE launchers (lunar for example) over ground-based launchers (see the references).

---

3. The ”laser on time” is proportional to $\xi^{3/4}$, and thus the placement of the DE launcher must be appropriate (Lubin 2016).
2.4.3. Scaling with data rate $R$

For any probe mass $m$ the initial data rate $R$ can be increased relative to the assumptions of Figure 2 by an increase in the receive aperture area $A_R$. For each 10x increase in $R$ (which requires a 10x increase in $A_R$) there is a 10x increase in data volume $V$ without an increase in latency $T_l$ (see Appendix A). Conversely the rate can be decreased with the benefit of a smaller receive aperture.

Another opportunity not accounted for in Figure 2 is an expectation that both technology and budgets can be expected to advance and expand during a decades-long mission. Probes can therefore be launched with an assumption of increasing data rate $R$ based upon planned infrastructure upgrades (such as larger receive aperture $A_R$) over the multi-decade life of the mission.

2.4.4. Refinements in the volume-latency

Two related secondary factors are unaccounted for in Figure 2. First is a decreasing RTG electrical power and data rate with time. Second is the opportunity to compensate for this by keeping peak power $P_p$ fixed. This increases PAR and hence BPP (as predicted by (3)). These refinements are quantified in Appendix A.2.

2.5. Outage mitigation

At optical wavelengths, atmospheric events (turbulence, water vapor, clouds, etc) impair (and even completely block) reception at random times, and the scattering of sunlight (including in the receive optics) may force reception to be limited to nighttime. Since the transmitter has no knowledge of when weather events occur, they have to be treated as random outages.

The siting of the receiver will introduce another source of outages if its latitude results in Proxima Centauri not being above the horizon at all times. A continuous view is possible if the receiver is sited at a sufficiently southerly latitude$^4$. Standard techniques for mitigation of outages include:

- **Error control coding (ECC)** adds redundancy to the scientific data, which can be used to reconstruct data lost to outages or other impairments (like quantum photon statistics).
- This ECC is more efficient when outage losses are spread randomly in time rather than clumped together. To this end, **interleaving** performs a pseudo-random permutation of the data. The effect of this is to spread the data from an individual image over time, and results in multiple images being transmitted concurrently. When interleaving is reversed in the receiver to recover the original ordering of the data, the outage losses are spread out in time and pseudo-randomized.

These techniques are illustrated in Section 4.4: There are two primary consequences of outage mitigation:

- A price has to be paid for outages, either in data rate, reliability objective, or signal power level. In Section 4.4 it is assumed that data rate and reliability are fixed, and receive-aperture area (and hence receive power) is increased to compensate for outages. For example, if the worst-case outage probability is $0.5$, then $A_R$ is slightly more than doubled.

- Interleaving introduces additional latency. Following interleaving, a single image will be spread over a long enough period (e.g. months or years) to statistically average outage events. Shorter periods could be accommodated, but there would be greater statistical variation and hence the penalty in receive aperture area increases. Averaging over multiple outage events is another compelling reason to prefer longer overall transmission times $T_d$.

2.6. Multiplexing multiple probes

If probes are launched once per week and the transmission time is $T_d = 20$ yr, the number of probes transmitting concurrently is $J_p = 1043$. Reducing $T_d$ or increasing the inter-launch interval beneficially reduces $J_p$.

Multiple probes complicate the downlink design because their communication has to be separated (whether there is a single or multiple receivers). This is called **multiplexing**, and there are some common alternatives:

- In **spatial multiplexing**, there are multiple receive apertures (or multiple beams from a single aperture). Each beam is designed to reject the signals from all probes save one. This is unlikely to be possible for the case of $J_p$ probes following similar nominal trajectories.

- In **wavelength-division** multiplexing (WDM) each probe is assigned a different wavelength at the receiver, so the signals can be separated by optical bandpass filters followed by a dedicated detector.

- In **time-division** multiplexing (TDM) each probe is assigned a different time slot at the receiver on a round-robin basis, so the probes' signals are separated in spite of wavelength-overlap. In this case there is a single optical bandpass filter and detector (although some wavelength agility may be necessary).

---

$^4$ The declination of Proxima Centauri is -62.67 degrees, so a continuous view is achieved by choosing a receiver site with a latitude more southerly than $(90 - 62.67) = 27.33$ degrees south. However, to avoid significant degradations in link quality when Proxima Centauri is low on the horizon due to increased atmospheric degradation effects, the receiver site should be more southerly than approximately 35 degrees south.
• In code-division multiplexing (CDM) the probes are assigned mutually orthogonal spreading sequences, which allows their signals to be separated (by cross-correlation with the spreading sequences) in spite of wavelength- and time-overlap.

For equivalent per-probe data rates $R$, to first order WDM, TDM, and CDM all expand the optical bandwidth by a factor of $J_p$. In the case of TDM this is because each probe has to transmit at rate $J_p R$ during its assigned time slot, and for CDM the spreading sequence expands the bandwidth of each probe’s signal by a factor of $J_p$. In practice the bandwidth is larger after accounting for guard bands (WDM) and guard times (TDM) due to imprecise receive wavelength and time slot management.

All these multiplexing methods have identifiable disadvantages. WDM requires precise transmitter wavelength control (a challenge complicated by Doppler shifts, see Section 2.8) and a parallel bank of receive-side optical bandpass filters to separate signals. TDM requires precise transmitter knowledge of Earth date/time and propagation distance, and consumes highly variable electrical power (which implies energy storage with an always-on electrical generator such as RTG). The performance of CDM in the context of low-power optical signals (and in particular its compatibility with high photon efficiency BPP) is unknown.

2.7. Bandwidth

A distinction should be made between two measures of bandwidth. The communication bandwidth $W$ represents the ideal bandwidth of the signal waveform input to the transmit laser (Section 3). The measurement bandwidth $B$ represents the bandwidth of the optical signal input to the detector, and governs the total power of the background radiation accompanying the received signal (Section 5). Obviously $B > W$, where $B$ has to be large enough to capture the entire transmitter intensity including a non-zero laser linewidth,\(^5\) any shortcomings in the receive-side optical bandpass filtering, and any increase in bandwidth required to accommodate uncertainty in the received signal wavelength.

When background is insignificant as assumed in this paper, the excess of $B$ over $W$ is not of concern, except for the rejection of $J_p - 1$ probe signals in the case of WDM. However, as background becomes a factor the design goal becomes to decrease $B$ to the order of $W$.

2.8. Doppler shift

Although the speed of each probe is expected to be constant once the launch phase is complete, there will be uncertainty in the speed of each probe and thus a difference in speed among probes. Correspondingly this results in an uncertainty and differentiation in received wavelength due to Doppler. This is relevant to several aspects of the design, including the receive aperture, background radiation, and multiplexing.

For transmit:receive frequencies $\nu_T: \nu_R$, the relativistic downlink Doppler shift obeys a product law $\nu_R = \nu_T \zeta_R \nu_T$ where $\zeta_T$ is attributable to transmitter motion and $\zeta_R$ is attributable to receiver motion (relative to a common notional inertial frame) (Messerschmitt 2017). The factor $\zeta_R$ has less impact because it is shared among all downlinks, it can be dynamically compensated at the receiver based on knowledge of Earth’s motion, and because $\zeta_R \approx 1$.

2.8.1. Transmitter red shift

Assume an inertial frame anchored to the Sun’s center (neglecting gravitational forces on the Sun). If the probe speed is $u$ away from Earth relative to this frame, the transmitter factor is $\zeta_T = \sqrt{(c - u)/(c + u)}$. For the assumed $u = u_0 = 0.2c$ this factor is $\zeta_T = 0.816$, indicating a 22.5% red shift.\(^6\) To avoid a reduction in receive aperture gain, the transmitter should compensate by introducing a countervailing blue shift. For example, if a receive wavelength of $\lambda_{R,0} = 1$ micron is desired, the transmit wavelength should be $\lambda_{T,0} = 816$ nm.

2.8.2. Uncertainty in probe speed

We can relate any uncertainty in probe speed $u$ to an uncertainty in frequency $\nu_R$. Assuming transmit wavelength compensation, the relative frequency tolerance is

$$\frac{\nu_R}{\nu_{R,0}} - 1 = \sqrt{c - u/c + u} - 1 \approx -1.042\frac{u - u_0}{c} + \ldots$$

Thus at $u_0 = 0.2c$ the fractional frequency tolerance and speed tolerance are related by

$$\frac{\Delta \nu_R}{\nu_{R,0}} \approx -0.208 \cdot \frac{\Delta u}{u_0}. \hspace{1cm} (4)$$

At wavelength $\lambda_0 = 1$ micron and speed tolerance $\Delta u / u_0 = \pm 1\%$, the variation in frequency is $\Delta \nu_R = \mp 625$ GHz.

2.8.3. Time dilation influence on data rate

Due to the increasing propagation time, the data rate observed by the receiver is smaller than the data rate generated at the transmitter. This reduction is 22.5% at the highest speed $u_0$, with a factor that becomes less significant as mass ratio $\xi$ increases. The data volumes $\mathcal{V}$ in Figure 2 account for this relativistic effect.

\(^5\) Today’s single-chip lasers have a linewidth of about 100 kHz, but in the future it may be possible to reduce this to order of 1 kHz.

\(^6\) Note that different $\nu_T: \nu_R$ are neglected in (1), which assumes the same wavelength for transmit and receive apertures.
2.8.4. Implications

The viable bandwidth of all multiplexed downlink signals will be limited by issues in implementing frequency-agile optical bandpass filters, and even more severely by the limited bandwidth that can be achieved in aperture synthesis. It is unlikely that probe speed can be controlled with sufficient precision (of the order of one part in $10^6$ to $10^9$) to accommodate these limitations. In that case, it will be necessary to compensate for variations in probe speed by active configuration of transmit wavelength. There are at least two possible methods:

- Autonomous configuration of transmit wavelength could be based on precise measurement of time to reach target, combined with precise knowledge of distance to target. Since the transmission of scientific data follows encounter, transmit wavelength configuration can precede downlink operation.

- If there is an Earth-to-probe uplink communication capability, a compensation factor based on measurement of Doppler shift at the receiver can be transmitted to the probe. This possibility is discussed further below.

2.9. Receive aperture area

The receiver consists of a large aperture followed by optical detection, the output of which is a stored record of the sequence of detected photons. This record becomes the primary mission archive, which can be processed off-line in non-real-time to extract the embedded scientific data. Thus there is little concern with receive processing intensity, and fortunately the major burden in the coding layer (Section 3) is at the receive end. The design of the receive aperture optics and photonics, on the other hand, raises difficult technological issues.

2.9.1. Aperture area

The baseline receive aperture size is determined by (1) and (2). Neglecting the Doppler-induced difference in transmit-receive wavelengths, and choosing the nominal values $d = D_t = 4.24$ ly, $R = 1$ bit per second, $BPP = 1$ bit per photon, $P_A = 10$ mW, $\lambda = 1$ $\mu$m, and $\eta = 1$, the resulting aperture area-product is $A_T A_R = 319.6$ cm$^2$k$m^2$. For $D_T = 10$ cm and $A_T = \pi D_T^2/4$, the resulting receive aperture diameter is $D_R = \sqrt{\pi A_R/4} = 2.276$ km.

Scaling from this value for any of the parameters is simple. An efficiency of $\eta = 0.5$ increases $D_R$ by $\sqrt{2}$ to $D_R = 3.219$ km. Any of the following changes reduces the diameter by $\sqrt{10}$ to $D_R = 720$ m:

- Reduce the data rate to $R = 0.1$ bps, or
- Increase the photon efficiency to $BPP = 10$ bits per photon (which is the objective pursued in Section 3), or
- Increase the transmit power to $P_A = 100$ mW, or
- Increase the transmit aperture diameter to $D_T = 10\sqrt{10} = 31.6$ cm.

This illustrates the substantial reduction in receive aperture area and diameter available by reducing data rate or increasing photon efficiency.

2.9.2. Factors increasing aperture area-product

A number of other environmental factors not taken into account in (1) will require a compensating increase in the aperture area-product. These include transmit aperture pointing error (related to attitude control of the probe), sources of absorption and scattering (not addressed in this paper), and sources of background radiation (Section 5).

2.10. Receive aperture field of view (FOV)

A receive aperture of the size envisioned here will necessarily be composed of $N > 1$ smaller sub-apertures, each with area $A_R/N$. Assume the $N$ sub-apertures are not mutually synthesized, and ask what requirements apply to $N$. Two performance issues in receive aperture design are the received signal power at detector output (related to total area $A_R$) and the FOV (related to sub-aperture area $A_R/N$).

2.10.1. Synthesis and FOV

If the sub-apertures are fully synthesized and diffraction-limited, the FOV of the overall aperture equals the FOV of the sub-apertures. For circular sub-apertures, the FOV can be approximated by the standard formula for the diameter of the Airy disk. Relevant to this question is the FOV, for which a useful reference is the diameter of the Airy disk corresponding to a range of angles $\Delta \theta$ rad,

$$\Delta \theta \approx 2.44 \cdot \lambda_0 \sqrt{\frac{\pi N}{4A_R}}.$$  

As expected the FOV increases as $\sqrt{N}$, and small FOV requires small $N$ (synthesis over larger sub-apertures).

2.10.2. Single probe tracking

It is difficult for the probe transmitter to outshine the entirety of emissions from the target star, even within a narrow optical bandwidth. This could be accomplished with sufficiently high transmit power or a large transmit aperture area (Section 5.2.1). Within the assumptions of this paper, negligible background radiation requires rejection of target star emission by the receive aperture. In other words, the FOV should exclude the target star.

The most extreme differences in probe trajectory are illustrated in Figure 1. The resolution of the probe from target star requires not only a small FOV but also accurate receive-aperture tracking of the probe trajectory to maintain the probe within that FOV. For example, the
radius of the exoplanet Proxima b orbit about Proxima Centauri is about 0.05 AU at a distance of 4.24 AU, corresponding to an angular separation of $\Delta \theta \approx 0.038^\circ$. Not only might the bullseye be considerably larger than this, but the parallax (due to a 2 AU baseline variation) results in $\Delta \theta \approx 1.54^\circ$. The effects of the non-zero bullseye and parallax are magnified during the post-encounter downlink operation, since the probe trajectories tend to diverge from one another.

2.10.3. Multiple probe tracking

Supporting multiple probe downlinks with a shared receive aperture and receiver is challenging due to the differences in probe trajectories. One option is to choose a single FOV sufficiently large to access all probes, which will include the target star within the FOV. The resulting substantial background (see Section 5.2.1) would necessitate a substantial increase in transmit power (to render background negligible in comparison) or reduction in photon efficiency BPP (due to the deleterious effects of background). A short visible wavelength $\lambda_0$ is also advantageous in minimizing this background (Section 5).

Negligible background can be preserved by attenuating target star emission, even while supporting downlinks from multiple probes. Some approaches that might be pursued include:

- Use TDM, with the receive aperture tracking a single probe during its assigned reception time window. In the extreme, use a reception time window sufficiently long to accommodate the data download from one probe in its entirety.

- Duplicate receive apertures and detectors, with each assigned to tracking a single probe.

- For a common optics, perform multiple synthesis and detection operations, with each assigned to tracking a single probe. (Such an approach is being explored at millimeter wavelengths for application to 5G cellular (Roh et al. 2014).) To avoid a reduction in SDR, do so with high quantum efficiency and without amplification.

- Use techniques similar to those being pursued in the direct imaging of exoplanets, which also requires attenuation of target-star radiation. These techniques include coronagraphs and interferometers (Traub & Oppenheimer 2010) and external occulters (also known as starshades) (Cash 2011; Lo et al. 2007). For example, a notch (high attenuation) permanently located at the target star location might be added to the aperture synthesis (this is illustrated in Appendix B.2 for a simple two-aperture synthesis).

A promising direction is to build on advances in direct exoplanet imaging. However, some differences inherent to the downlink design challenge should be noted:

- The observation platform is ground-based rather than space-based. Among other issues, spacecraft attitude control is replaced by dynamic pointing in aperture synthesis and atmospheric turbulence becomes an issue.

- The total aperture area is much larger due to the weak source (an RTG powered transmitter rather than whole-planet starlight reflection).

- The optical bandwidth is narrow and the wavelength can be chosen for engineering rather than scientific objectives. It is estimated in Section 5.2.1 that the degree of starlight attenuation required to eliminate background as a consideration is, for Proxima Centauri, a relatively modest 1300x at 400 nm, but increases to $8.1 \cdot 10^5$ at 1 $\mu$m.

- While the direct imaging observation of an exoplanet may be relatively short-lived, a probe downlink may operate for decades to a century or longer.

2.11. Bandwidth limits in aperture synthesis

The sub-apertures are likely to be implemented as an interferometer, synthesized from an array of smaller elements. The preceding analysis assumes a perfectly-packed array, and the beam sidelobes will increase as the array dimension is increased due to lower packing density.

2.11.1. How it works

Each element provides a relatively large field of view, and thus can be pointed with relaxed accuracy (Burke & Graham-Smith 2009). If the probe is at a vertical position relative to the plane of the array, the optical outputs can simply be combined. However, if the probe is at some angle $\theta$ to the vertical, there are geometric delays $\tau_g$ introduced between elements due to the differing path lengths from source to element, which must be compensated by instrumental delays $\tau_i$ at the output of the apertures with the shorter paths. The geometric delay $\tau_g = b \sin(\theta/c)$ for baseline distance $b$ is angle-dependent. The instrumental delay is thus also angle-dependent, and determines the direction of the main lobe.

2.11.2. Bandwidth limitations

Significantly, aperture synthesis is effective over a limited bandwidth, and thus interacts with the multiplexing method and variations in frequency due to Doppler shifts. A two-element synthesis is analyzed in Appendix B.2. For a two-element interferometer, the baseband bandwidth $B$ is determined by $B(\tau_g - \tau_i) \ll 1$. (This is shown in Appendix B.2 for a single modulation-coded ON-OFF signal as described in Section 3.) Thus, at the angle $\theta$ precisely matched by the delay (where $\tau_g = \tau_i$)
the bandwidth is large. However, if there is any delay-matching error at the angle of the actual point source, or if the source has a finite size (so that delay matching cannot be perfect across the entire source), the bandwidth becomes restricted due to $\tau_g \neq \tau_i$.

2.11.3. Aperture design challenges

There are at least two. First is implementation of accurate delay matching, which requires instrumental delays precisely controlled in the optical domain. For the three synthesized sub-aperture diameters \{20cm, 60cm, 30m\} the maximum delays are \{0.67ns, 2ns, 100ns\}. The second challenge is the finite bandwidth supported by the synthesis. In the absence of instrumental delay, the smallest bandwidth in the three cases is on the order of \{1.5 GHz, 500 MHz, 10 MHz\}. Dividing the bandwidth among $J_p = 1000$ probes with WDM multiplexing yields a per-probe bandwidth (in the first two cases) of \{1.5 MHz, 500 kHz\}. The bandwidth requirement for the data signal is considered further in Section 3, but is considerably smaller than 1 MHz. The larger FOV required for multiple probes is consistent with the larger bandwidth requirement from multiplexing those probes.

2.12. Bandpass filtering

Bandpass filtering in the optical domain is needed for two reasons:

1. Elimination of out-of-band background radiation. The main reason background radiation may be negligible (per the assumptions of this paper) is the low data rates that are typical due to transmit power limitations and great propagation distances. Specific bandwidths will be justified in Section 3.

2. Separation of multiplexed signals from multiple probes using WDM.

Cavity-based optical bandpass filters with high selectivity have been demonstrated (Spencer et al. 2014, 2012). With configuration of transmit wavelength at sufficient precision, it may be possible to avoid frequency agility in these bandpass filters, which would be difficult to achieve.

2.12.1. WDM

An additional requirement with WDM is for a bank of bandpass filters with hundreds or thousands of channels with high quantum efficiency and without amplification. An alternative is a detector technology which estimates the energy (e.g. wavelength) of each detected photon. The separation of out-of-band background radiation from signal and the separation of multiple WDM channels could then be performed in the post-detection processing.

2.13. Uplink for probe configuration

It was concluded in Section 2.8 that it is necessary to configure the transmit wavelength to adjust for Doppler. Fortunately, the speed is constant following the propulsion period, so a one-time configuration may suffice. It may also be necessary to perform a course correction following initial propulsion to achieve a small enough bullseye for scientific purposes.

These requirements suggest the need for a near-Earth communication uplink capability. Following the end of the propulsion phase, the probe could transmit a signal toward Earth, and a broadband receiver could measure both the actual trajectory of the probe and the signal wavelength. Course and wavelength correction factors can be transmitted back to the probe. Fortuitously this two-way communication occurs near to Earth, and thus suffers minimal attenuation. Background radiation should be less a concern, allowing the receive aperture to operate in non-synthesis mode with a large FOV. The uplink communication power can also be high, greatly simplifying the probe’s uplink receiver. Uplink operation time, like launch time, becomes a contributor to downlink outage time (Section 2.5).

2.14. Direct vs heterodyne detection

Heterodyne mixes the optical signal with a local oscillator (LO), and the optical detector square-law detector results in components of sum and difference frequencies (Appendix B.1) and also amplifies the signal. It can be arranged for the difference-frequency component to fall at radio frequencies. The previous discussion has assumed direct detection, which eliminates the LO and uses a detector capable of directly counting photons without amplification or frequency shift.

2.14.1. Advantages

Heterodyne has some compelling advantages. It provides amplification of the signal level at the detector output. Channel separation and bandpass filtering can be performed using available microwave technologies. In contrast to the intensity modulation that is compatible with direct detection, heterodyne detection admits the possibility of modulating phase as well as amplitude and thus opens up a wider class of techniques within the coding layer.

2.14.2. Disadvantage

Unfortunately, due to the shot-noise introduced by the large LO, heterodyne cannot achieve a comparable photon efficiency to photon counting. There is a hard limit of $BPP < \eta/\log 2 = 1.44\eta$ bits per average detected photon for a detector quantum efficiency $\eta$ (Gordon 1962). In contrast, there is no theoretical limit on BPP for direct detection other than our ability to implement high PAR (Appendix C.2). Approaching these respective heterodyne and direct-detection fundamental
limits requires comparable levels of processing in the coding layer, although the waveforms and algorithms will differ significantly.

Due to its greater potential photon efficiency, there is a substantial payoff for solving the technological challenges inherent in direct detection. For this reason, our focus is on direct detection in the remainder of this paper.

3. MODULATION-CODING LAYER

The communications portion of the probe transmitter and terrestrial receiver are organized into layers as illustrated in Figure 3. Most of the issues discussed in Section 2 fall into the physical layer, which includes everything from a waveform (voltage vs. time) at the input to detected photon events at the output. At the transmitter a laser converts the waveform into light intensity vs. time, and thus the scientific data is embedded in the intensity of the received optical signal. For convenience, we measure that intensity by the average rate $\Lambda(t)$, which for convenience we take to be $0 \leq \Lambda(t) \leq 0.1$. During interval $T_c$, at data rate $R$ the number of scientific data bits represented by this intensity is $RT_c$. These bits can be represented by transmitting one of $N_c$ intensity waveforms, where $N_c = 2^{RT_c}$. Let the rate of detected photons for the $n^{th}$ transmitted waveform be $\lambda_n(t)$ for $1 \leq n \leq N_c$. Then the entire set of such waveforms $\{\lambda_n(t), 1 \leq n \leq N_c, 0 \leq t \leq T_c\}$ is called a codebook, each element of this codebook is called a codeword, and the process of mapping $RT_c$ data bits into the corresponding codeword is called channel coding.

3.1. Channel coding for intensity modulation

A direct detection receiver estimates signal power vs. time, and thus the scientific data is embedded in the intensity of the received optical signal. For convenience, we measure that intensity by the average rate $\Lambda(t) \geq 0$ of detected photons vs. time, which is proportional to the instantaneous transmitted power. As now shown, the choice of intensity modulation waveforms can have a significant impact on the photon efficiency $BPP$. We now describe how $\Lambda(t)$ is related to the scientific data being communicated.

3.1.1. Coding

Consider a time interval with duration $T_c$, which for convenience we take to be $0 \leq t \leq T_c$. During interval $T_c$, at data rate $R$ the number of scientific data bits represented by this intensity is $RT_c$. These bits can be represented by transmitting one of $N_c$ intensity waveforms, where $N_c = 2^{RT_c}$. Let the rate of detected photons for the $n^{th}$ transmitted waveform be $\lambda_n(t)$ for $1 \leq n \leq N_c$. Then the entire set of such waveforms $\{\lambda_n(t), 1 \leq n \leq N_c, 0 \leq t \leq T_c\}$ is called a codebook, each element of this codebook is called a codeword, and the process of mapping $RT_c$ data bits into the corresponding codeword is called channel coding.

3.1.2. Decoding

At the receiver, the actual detected photons are examined to estimate which of the $N_c$ codewords most closely approximates the actual pattern of detected photons. In this channel decoding step, the resulting index $n$ into the codebook is mapped into the corresponding set of $RT_c$ data bits.
Because the pattern of detected photons is random (reflecting its quantum nature) this decoding does not always choose the correct codeword. For any choice of a codebook, the designer can utilize the Poisson statistics which model quantum photon counting to calculate some appropriate measure of error probability $P_e$, which is typically chosen to be bit or codeword error probability.

3.2. Fundamental limit

A coding theorem divides the possible values of $\{R, P_e\}$ into two regimes associated with a threshold on $R$ (Gallager 2008; Cover & Thomas 1991; Messerschmitt 2008). When $R < C$, where $C$ is called the channel capacity, any desired degree of reliability $P_e$ can be achieved in principle. Specifically, for any $\epsilon > 0$ there exists a codeword length $T_c$ sufficiently large and a corresponding codebook choice such that the resulting $P_e < \epsilon$. Generally when $\epsilon$ is chosen smaller, then $T_c$ must be chosen larger. Conversely when $R \geq C$, $P_e$ is always bounded away from zero for arbitrarily large $T_c$ and all possible codebook choices.

In everyday language, reliable communication is known to be feasible when $R < C$ and not feasible when $R \geq C$. The coding theorem does not tell us how to design a reliable codebook, but speaks only to feasibility.

3.2.1. Interpretation

Coding and decoding achieve improved reliability of data recovery (smaller $P_e$) through statistical averaging of the random photon detection events. It is advantageous to choose a larger $T_c$ because this results in a larger number of bits $RT_c$ being determined from a correspondingly larger number of photon detection events. Each and every decoded bit among the $RT_c$ bits is determined by the totality of detected photon events over time duration $T_c$.

Based on (2) since $A_A$ is a fixed property of the physical layer, it follows that any boundary on $R$ that can be achieved in the coding layer is effectively a boundary on the achievable photon efficiency BPP. When we say for example that BPP = 10, we mean specifically that arbitrarily small $P_e$ can be achieved in principle by choosing a sufficiently long codeword duration $T_c$ and an appropriate codebook choice, and that in the process $RT_c$ data bits can be recovered from an average of $RT_c/10$ average detected photons.

The achievable reliability improves with larger $RT_c$. Achieving a high BPP with reliable recovery of scientific data generally requires averaging over a large number of photon detection events. Thus $10^5$ bits conveyed by $10^4$ average detected photons can achieve a smaller $P_e$ than $10^4$ bits conveyed by $10^3$ average detected photons due to the longer averaging period. This assumes the choice of a meritorious codebook in each case.

3.2.2. Constraints

If we allow infinite transmit power, any data rate $R$ can be achieved reliably. To get practically meaningful results, we must place constraints on the transmit power, which correspond to constraints on the codewords $\Lambda_n(t)$. These are the peak $A_P$ and average $A_A$ power constraints

$$0 \leq \Lambda_n(t) \leq A_P, \ 0 \leq t \leq T_c$$

$$\frac{1}{T_c} \int_0^{T_c} \Lambda_n(t) \cdot dt = A_A, \ 1 \leq n \leq N_c.$$  

We make two simplifying assumptions, which it turns out have no effect on the capacity $C$. First, each and every codeword is constrained to have the same energy $T_c A_A$. Also, every codeword has an ON-OFF character ($\Lambda_n(t) = \Lambda_P$ or $\Lambda_n(t) = 0$ for $0 \leq t \leq T_c$). Photon rates $\{A_A, A_P\}$ are referenced to the detector output, taking account of everything that happens in the physical layer (including quantum efficiency).

3.2.3. Capacity for photon counting

For the given power constraints, and assuming the appropriate Poisson statistics for photon detection events, the capacity $C$ can be determined. The result is given in Appendix C for the general case where $A_B > 0$, including establishing the regime in which $A_B$ can be neglected. This regime is approximately $SBR > 10^2$, where the signal-to-background ratio is $SBR = A_A/A_B$.

When background is neglected by choosing $A_B \to 0$, the capacity becomes

$$\text{BPP} = \frac{C}{A_A} = \begin{cases} \text{PAR} \cdot \frac{\log_2 e}{e} & 1 \leq \text{PAR} \leq e \\ \frac{\log_2 \text{PAR}}{e} & \text{PAR} > e \end{cases}. \quad (5)$$

This reduces to (3) for the larger PAR of interest. This fundamental upper limit on BPP is plotted as the dashed line in Figure 4, where (5) tells us that any point below is a feasible operating point with any desired reliability objective. This bound is monotonically increasing with PAR, starting with BPP = 0.53 bits per photon at PAR = 1, reaching BPP = 1.44 at PAR = e. From there, as PAR $\to \infty$, BPP $\to \infty$ logarithmically in PAR. Although there appears to be no limit on the BPP that can be achieved, in fact it is limited by both practical and fundamental factors.

We conclude that there are two ways to increase the data rate $R$:

- Increase the average rate of detected photons $A_A$, which can be accomplished with higher average power, PAR, or a longer time duration $T_c$.

7 The uncertainty principle of quantum mechanics limits the PAR that can be achieved. Mundane practical concerns like the feasible speed of the electronics generally intervene at much smaller levels of PAR (Butman et al. 1982).
Figure 4. The photon efficiency BPP as a function of logarithm of the peak-to-average ratio PAR, demonstrating the benefit of designing for a larger PAR. The dashed line is the upper limit for reliable data recovery given by (5). The single point labeled “a” is for uncoded OOK (Section 3.4), where BPP = 0.125. The points labeled “b” are for uncoded PPM with different values of PPM order $M = 2^m$ for $1 \leq m \leq 20$ (Section 3.5). The points labeled “c” denote the upper limit for reliable data recovery for PPM with different values of $M = 2^m$ for $1 \leq m \leq 20$ combined with another layer of channel coding (Section 3.6).

power $P_A$ in the transmitter or larger optics aperture area product $A_T A_R$. When $P_P$ remains constant, as we expect, the impact of increasing $P_A$ is reduced by the corresponding reduction in PAR, although there remains a net benefit.

- Increase the photon efficiency BPP by increasing PAR. Thus larger peak power $P_P$ (resulting in larger $\Lambda_P$) is always desirable.

3.2.4. Factors determining PAR

The peak power $P_P$ is an important parameter of the transmitter design. In the following we assume that $P_P$ is fixed over the entire mission, independent of $P_A$. There are physical limitations on the diode laser technology which limits the $\Lambda_P$. It is possible to “gang” multiple laser diodes transmitting simultaneously to overcome this limitation, although this results in more complex optics and greater mass.

3.2.5. Intuition

For ON-OFF codewords, the duty factor (fraction of time the signal is ON) is the reciprocal of PAR. Usually as PAR grows the ON cycles become shorter, raising the bandwidth. Thus PAR and bandwidth are intimately related. Due to shorter ON cycles, large PAR is attractive for conveying accurate timing information. Representing data by timing accuracy is attractive because timing does not consume energy (energy consumption reduces photon efficiency BPP). We will see a concrete example of this in the PPM example (Section 3.5).

3.3. Layered channel coding

Now we address the issue of how to achieve high BPP with concrete implementations. While the codebook as we have described is a good conceptual tool for understanding channel coding, it fails miserably in practice due to the exponential growth in codebook size as $T_c$ increases, because it has $2^{RT_c}$ codewords. This quickly results in unreasonable expectations for storage and also decoding processing (required to compare the detected photon pattern with all possible codewords). In practice coding/decoding must be performed algorithmically, thus substituting processing for storage resources. We now describe an algorithmic approach which can reasonably close to the fundamental limit of channel capacity.

3.3.1. Data reliability

We are now moving from the computation of channel capacity to consideration of concrete techniques designed to achieve high photon efficiency BPP. For any concrete implementation, an issue is always the achieved reliability as measured by an error probability $P_e$. Thus a given BPP is always associated with a reliability objective $P_e$, and generally speaking as we relax our reliability objective we can achieve higher BPP. The fundamental limit of (5) is a useful comparison, because it indicates how much greater a BPP may be possible with a more sophisticated codebook design.

3.3.2. Binary codewords

As mentioned earlier, we can assume that all codewords are ON-OFF without compromise in the achievable rate $R$.

3.3.3. Regularity in codewords

Define $T_s$ as the shortest ON time interval across all codewords. Then we impose a regularity constraint that all ON time durations are multiples of $T_s$, which is called the slot time. This regularity will reduce the data rate $R$ that can be achieved reliably. However, we demonstrate by example that the resulting penalty in BPP can be small.

---

8 Unnecessarily reducing $P_P$ would have a deleterious effect on the photon efficiency BPP that can be achieved.

9 Increasing PAR and hence bandwidth causes SBR to grow, and to enter the regime where background is significant. This is another reason (5) is invalidated for very large PAR.

10 For high reliability (small $P_e$) the resulting BPP will generally be smaller than (5). If the reliability is greatly relaxed (e.g. $P_e \approx 0.5$) achieving a BPP larger than (5) is not ruled out by the coding theorem. However, this is of little practical interest.
3.3.4. Average photons per slot and codeword

A quantity of considerable significance to reliability is the average number of detected photons during slot $T_s$, which we call $K_s$. This is related to the peak power through $K_s = T_s \Lambda_P$. If we think of the data being encoded through the timing of an ON pulse, then the reliability with which that data can be recovered increases with $K_s$. Quantum mechanics tells us that the actual number of detected photons $Y$ is a random variable obeying a Poisson distribution,

$$\text{Pr} \{ Y = k \} = \frac{K_s^k e^{-K_s}}{k!}. \quad (6)$$

For our purposes the only relevant case occurs when $K_s \neq 0$ and the probability of zero photons being detected is $\text{Pr} \{ Y = 0 \} = e^{-K_s}$.

3.3.5. Modulation code

Rather that perform channel coding all at once with a very large $T_c$, as described in Section 3.2.2, the channel coding is typically constructed in layers (two layers in Figure 3). The bottom layer is the modulation code, which acts as a sophisticated form of digital-to-analog-to-digital converter.

For the convenience of the higher layers the codewords are index by $m$ bits, so the number of codewords in the codebook is constrained to $M = 2^m$. In the transmitter, each codeword of duration $T_c$ indexed by $m$ bits is mapped into an ON-OFF light intensity waveform of duration $T_c$, where the minimum ON time is slot time $T_s$. Two modulation codes (OOK and PPM) are illustrated in Figure 5.

The receiver observes a set of photon detection events and chooses the codeword (among $M$ possibilities) that was most likely to have been transmitted. The $m$-bit index of that codeword is the output, which is passed to the next layer.\footnote{Often more information quantifying the reliability of that decision is added. For example, we will see in the case of PPM that it is advantageous to add another bit which signals an erasure.}

We now consider two specific modulation code layer designs of Figure 5.

3.4. On-off keying (OOK)

OOK is the simplest possible modulation code, and thus provides a baseline for gauging the impact of using more sophisticated techniques. As illustrated in Figure 5a, it transmits a sequence of ON-OFF pulses, each with duration equal to the timeslot $T_s = T_c$. There are only $M = 2$ codewords in the codebook, with $m = 1$ bit in and one bit out.

Suppose the input for the $k$th modulation codeword is $X_k$ and the output after detection of that codeword is $Y_k$. Thus for OOK $X_k \in \{0, 1\}$, where $X_k = 0$ is transmitted as an OFF pulse and $X_k = 1$ is transmitted as an ON pulse. The average number of photons detected at the receiver for an ON pulse is the parameter $K_s$. The peak power at the detector is thus $\Lambda_P = K_s/T_s$.

3.4.1. Reliability

We are concerned with the data rate $R$ that can be achieved with OOK for a specified reliability objective. Assume that the following detection scheme is utilized at the receiver. If zero photons are detected during slot $k$, the output is declared to be $Y_k = 0$, and if one or more photons are detected $Y_k = 1$. We build a statistical model for this binary in-out channel by noting the transition probabilities. These are drawn from (6),

$$\begin{align*}
\text{Pr} \{ Y_k = 0 | X_k = 0 \} &= 1 \\
\text{Pr} \{ Y_k = 1 | X_k = 0 \} &= 0 \\
\text{Pr} \{ Y_k = 0 | X_k = 1 \} &= e^{-K_s} \\
\text{Pr} \{ Y_k = 1 | X_k = 1 \} &= 1 - e^{-K_s}
\end{align*} \quad (7)$$

3.4.2. Without ECC

The case where the ECC layer is omitted entirely offers a baseline for comparison with more sophisticated techniques. Assume that $X_k = \pm 1$ are equally likely, so $P_e = e^{-K_s}/2$, and let our reliability goal be $P_e = 10^{-7}$. Then we require $K_s = 15.4$. The average number of photons per slot is $K_s/2$ and the number of bits conveyed per slot is one, so that $BPP = 0.13$.

To compare OOK with the fundamental limit of (5), for equally likely $X_k = \pm 1$ the peak-to-average ratio of OOK is PAR = 2. The performance of OOK without ECC is the point labeled “a” in Figure 4a. This quantifies what could theoretically be accomplished (for equivalent PAR) by adding the ECC layer to OOK, which is an increase in BPP by a factor of $1/0.13 = 7.7$.

A much bigger opportunity evident from Figure 4 is the much larger gains in BPP that could theoretically be
obtained by choosing a modulation code that achieves a larger PAR.\textsuperscript{12} This is the idea behind PPM, which is considered next.

3.5. Pulse-position modulation (PPM)

PPM is illustrated in Figure 5b. It has an intrinsically large PAR, combined with other desirable properties. One PPM frame of duration $MT_s$ is composed of $M$ slots of duration $T_s$. The constraint is introduced that there is exactly one ON slot and $(M-1)$ OFF slots. Since there are $M$ alternative locations of the single ON pulse, the number of bits conveyed by one PPM frame is $m = \log_2 M$. Further PAR = $M$, desirably without any dependency on the statistics of the data being transmitted. Typically we choose $M = 2^m$ so that each PPM frame represents $m$ data bits.

3.5.1. Reliability

As before let $K_s$ be the average number of photons detected in the single ON slot, and of course there are 0 average photons detected in the $M-1$ OFF slots. Since there is only a single ON slot, $K_s$ is also the average detected photons per PPM frame. Assume that the modulation decoding layer in the receiver counts the number of photons detected in each of the $M$ slots within an individual PPM frame. When background radiation is neglected, the result will be zero photons detected in the OFF timeslots, and any number of photons (including zero) in the single ON timeslot. The only type of error is thus an erasure, which occurs when there are no detected photons in the ON slot, and hence in any slot. Desirably the decoder can flag this erasure, but it cannot infer anything about the $\log_2 M$ bits that are represented by this erased PPM frame.

Suppose the input to the $k$th PPM frame denoted by $X_k$ and the corresponding PPM decoder output is $Y_k$. Then the input-output relationship looks like

$$X_k \in \{1, 2, \ldots, M\} \rightarrow Y \in \{1, 2, \ldots, M, \mathcal{E}\}. \quad (8)$$

Each input $X_k$ (for the $k$th PPM frame) tells PPM which timeslot is ON, and the decoder by observing detected photons can determine which timeslot was ON, with the additional possibility of an erasure (which is labeled as output $\mathcal{E}$). An erasure occurs with probability $e^{-K_s}$. Note that an erasure is spelled out by the special symbol $\mathcal{E}$ because this frame is known to be in error. This flags corrupted data, which proves useful in the ECC layer.

We can define the reliability by the probability of an erasure $P_e = e^{-K_s}$. For example, $P_e = 10^{-7}$ requires $K_s = 16$. In this case $m$ bits are conveyed by 16 photons, and thus $\text{BPP} = m/16$. The resulting BPP is plotted in Figure 4 by the points labeled “b”. Unlike OOK, PPM provides a straightforward way to achieve high PAR and, associated with this a higher BPP. However, it only achieves 6.25% of the fundamental limit for the same PAR and with this reliability. In other words, it gives up 93.75% of the feasible gain in photon efficiency for equivalent PAR.

3.5.2. Interpretation

PPM by itself, without the benefit of an ECC layer, is actually a channel coding scheme with a codeword duration $T_c = MT_s$. Although its photon efficiency BPP falls considerably short of the fundamental limit for the same PAR, unlike OOK it does benefit from a larger PAR. One explanation for this benefit is that PPM conveys $m$ bits of data through the timing or location of a single ON pulse. Using timing to represent data bits comes for free in terms of energy consumption.

PPM frames interpreted as codewords display redundancy. This means that of the possible codeword combinations in $M$ slots, only a small subset are chosen to represent the data. By allowing all $M$ slots to independently be ON or OFF, $M$ data bits would be represented by $2^M$ different codewords. In fact for PPM only $\log_2 M$ data bits are represented by $M$ different codewords.

Redundancy is a general theme in channel coding, and we will see it more explicitly realized in the ECC layer. The motivation for redundancy is to improve reliability by making it less likely that quantum fluctuations result in an incorrect decoding by keeping the codewords more sparsely located.

3.5.3. Bandwidth

In the absence of ECC, PPM requires greater bandwidth than OOK for the same data rate $R$. For OOK we get $R = 1/T_s$ (one bit of data is conveyed in time $T_s$) and for PPM $R = m/(MT_s)$ ($m$ bits of data are conveyed during a time $MT_s$). Thus for equivalent data rate $R$, the slot $T_s$ for PPM has to be a factor of $m/M$ smaller, implying that the bandwidth $W$ of the data signal will be a factor $M/m$ higher.

This bandwidth expansion factor, as it turns out, is directly related to the improvement in photon efficiency BPP. As $M$ grows the bandwidth $W$ grows as BPP grows with it.

3.6. Fundamental limit revisited: PPM with ECC

We have seen that PPM by itself falls considerably short of the theoretical photon efficiency BPP given by (5) for the same PAR. Is it possible for an ECC layer to recover this entire gap? Although the answer is no, it is possible to come close with an advantageous choice of $K_s$. To see this, we now revisit the channel capacity question. A concrete example of increasing BPP through ECC is deferred to Section 4.
3.6.1. Longer codewords

The gap between PPM and the fundamental limit is explained in large part by the relatively short-duration codewords in the codebook. The channel coding theorem (Section 3.2) suggests that increasing the length of the codewords is a key to higher data rates and hence higher photon efficiency. This could be accomplished by increasing the PPM order \( M \), but that would have the side effect of increasing \( \text{PAR} = M \). A fairer comparison is to hold \( \text{PAR} = M \) fixed, because that is determined by the technology and design of the physical layer.

A different way to form longer codewords is by grouping together \( L \geq 2 \) PPM frames to form an effectively longer codeword of length \( T_c = L M T_s \). These longer codewords create a codebook with \( 2^{LM} \) codewords, each such codeword indexed by \( LM \) bits. The transmitted light intensity waveform would then consist of \( L \) concatenated PPM frames, each with its single ON pulse in a potentially different location within its respective frame. As desired, every codeword would still be constrained to have \( \text{PAR} = M \).

3.6.2. Equivalent discrete input-output channel

Relevant to this concatenated codeword scheme, there are two related questions:

- What is the theoretical upper limit on the BPP that can be achieved with reliable data recovery? How does this compare to (5)?
- How would we actually construct a codebook consisting of \( L \) PPM concatenated codewords, and what would be the resulting BPP?

We address the first question now, and the second question is deferred to Section 4.

3.7. A return to the fundamental limit

Consider the discrete-time channel defined by (8). Associated with this definition is a set of transition probabilities that are a function of the parameters \( M \) and \( K_s \). This is in fact the channel that is seen by the ECC layer in Figure 3. The ability of the ECC layer to achieve a given bit rate at a given level of reliability, assuming a PPM modulation code, is thus completely determined by this channel with its input-output alphabet and its transition probabilities.

Just like the underlying photon-counting channel that led to the capacity relation (5), the capacity of the channel defined by (8) can be determined. Call the capacity of this channel as \( C_{\text{PPM}} \). The theoretical limit on the photon efficiency BPP of this channel is known to be (Dolinar et al. 2011)

\[
\text{BPP}_{\text{PPM}} = \frac{C_{\text{PPM}}}{\Lambda_A} = \gamma_{\text{PPM}} \cdot \log_2 M
\]

\[
\gamma_{\text{PPM}} = 1 - e^{-K_s} \frac{1}{K_s}.
\]

This has the same coding theorem interpretation. That is, as long as the information rate satisfies \( R < C_{\text{PPM}} \), and only then, arbitrary reliability can be achieved in theory through a concrete design of a codebook. This requires that \( L \) be as large as necessary to achieve the desired reliability.

3.7.1. Remaining gap to the fundamental limit

Interestingly (9) has the same form as (5) for equivalent \( \text{PAR} = M \), except for the \( \gamma_{\text{PPM}} \) term. This term is in the range of \( 0 < \gamma_{\text{PPM}} < 1 \), so unsurprisingly \( \text{BPP}_{\text{PPM}} < \text{BPP} \). Further, \( \gamma_{\text{PPM}} \) increases monotonically as \( K_s \) decreases, with an asymptote of \( \gamma_{\text{PPM}} \to 1 \) as \( K_s \to 0 \). This demonstrates that it is feasible for an ECC layer, in conjunction with a PPM modulation code, to approach arbitrarily close to the fundamental limit of (3) for equivalent PAR.

3.7.2. Choice of \( K_s \)

To achieve \( \gamma_{\text{PPM}} \approx 1 \), parameter \( K_s \) should be very small. However there are practical limits on how small \( K_s \) can be chosen. One is the bandwidth \( W \) of the signal. Recall that \( K_s = T_s \Lambda_P \) where \( \Lambda_P \) is fixed by the transmit peak power \( P_P \) and fixed parameters such as distance and aperture size. As \( K_s \to 0 \) it follows that \( T_s \to 0 \), which implies that bandwidth \( W \to \infty \). Eventually we will violate the uncertainty principle, but long before that the short pulses are difficult to generate with the current electronics technology. Also increasing bandwidth \( W \) will eventually draw us into the regime of material background radiation, violating the conditions under which (9) is valid.\(^{13}\)

3.7.3. Example

Suppose we seek a data rate close to \( R = 1 \) bits per second with a photon efficiency close to \( \text{BPP} = 10 \) bits per photon by starting with an average received power \( \Lambda_A = 0.1 \) photons per second and a PPM modulation

| Parameter                     | Value  |
|-------------------------------|--------|
| Average power \( \Lambda_A \) (photons per s) | 0.1    |
| Photons per frame \( K_s \)     | 0.1    |
| PPM frame rate (s\(^{-1}\))    | 1      |
| Capacity \( C_{\text{PPM}} \) (bps) | 0.952  |
| Data bits per frame @ capacity | 0.995  |
| Frame erasure probability \( P_e \) | 0.905  |
| Slot time \( T_s \) (\(\mu s\)) | 976    |
| Bandwidth \( W \) (kHz)       | 1.024  |

\(^{13}\) Another consideration is the visibility of PPM frames to the receiver, which is necessary for derivation of the timing of the PPM frames. In practical terms, a certain fraction of frames have to be visible for the receiver to be able to estimate the current rate and timing of PPM frames.
code with $M = 2^{10} = 1024$. This requires that the physical layer be capable of PAR = 1024. Then the consequences of choosing two small values $K_s = 0.1$ or 0.01 photons per frame are shown in Table 5.

Notably the rate at which frames are generated increases as $K_s$ decreases. At $K_s = 0.1$ the capacity $C_{PPM}$ falls 4.8% short of our ideal of 1 bps, and this is reduced to a 0.5% shortfall at $K_s = 0.01$. The data bits conveyed per frame at capacity decreases as $K_s$ decreases because of the increased frame rate.

3.7.4. Redundancy

Perhaps the most surprising line in Table 5 is the average fraction of frames that are erased (zero photons are detected), which increases from 90.4% to 99%. That implies that only 9.6% or 1% of the PPM frames on average are visible to the receiver. This PPM modulation code with small values of $K_s$ is, in isolation, extremely unreliable. The price paid for achieving high photon efficiency BPP is to starve the PPM modulation code of photons, driving it into a very unreliable regime. The savior is the ECC layer, which has an opportunity (and (9) tells us it has the theoretical ability) to overcome the overwhelming number of erasures and reliably recover our scientific data bits.

The basis for this ECC opportunity is the large redundancy available. Each PPM frame represents only 0.952 or 0.995 bits, but each frame conveys $2^{10}$ possibilities, which could represent 10 bits. Thus the available redundancy is 9.048 or 9.005 bits per PPM frame. What (9) tells us is that this is the minimum redundancy required at the fundamental limit for these values of $K_s$.

3.7.5. PPM bandwidth

From the relation $K_s = T_s A_P$ we find that the bandwidth of the PPM for our choice of $K_s$ is $W \approx 1/T_s = \Lambda_A/K_s$. Thus as $K_s$ decreases, the bandwidth of optical signal increases. For the example in Table 5 these values are about 1 and 10 kHz, compared to a data rate of 1 bps. Generally we find that $W \gg R$, or the bandwidth of our optical signal is far greater than the data rate that it supports. For fixed $R$, $W$ increases if $K_s$ is smaller (resulting in a smaller gap to the fundamental limit) or if BPP is larger (because we have chosen a larger $M$). Physically all this occurs because the PPM frames have shorter duration (a higher rate of PPM frames).

3.7.6. Scaling data rate

While the parameters of Table 5 seek a data rate $R \approx 1$ bps, this rate can be increased to 10 bps with one of two modifications to the physical layer:

- Increase PAR = $M$ in the physical layer by a factor of 1024 to $M = 2^{20} \approx 10^6$ through a commensurate increase in peak power $P_P$. This will reduce $T_s$ and increase bandwidth $B$ by the same factor, and thus increase background and reduce SBR.
- Increase signal power $\Lambda_A$ by a factor of 10 through a commensurate increase in transmit power or transmit-receive aperture area product. This will have the side effect of increasing the PPM frame rate, bandwidth $W$, and background radiation $\Lambda_B$ by the same factor. However, SBR is not affected.

3.7.7. Interpretation

For fixed data rate $R$, seeking out high BPP requires an increase in the PPM frame rate. This in turn shrinks the slot time $T_s$, resulting in fewer detected photons per slot $K_s$. Smaller $T_s$ increases the optical bandwidth, with the side effect that background power $\Lambda_B$ is increased and SBR is decreased. The reliability with which data is extracted from the PPM frames is reduced, but this can be overcome in principle in the ECC layer.

Why is there an advantage in using small $K_s$? For a fixed received average power $\Lambda_A$, smaller $K_s$ increases the PPM frame rate, reduces the PPM ON pulse duration, and increases the bandwidth. The fundamental idea in achieving high BBP is to employ timing precision to communicate data. If PAR is fixed, then smaller $K_s$ results in shorter-duration PPM frames, which in turn makes the ON pulses shorter, and thus increases the timing precision. Those pulses are less reliably detected, but those that are detected have greater timing precision. The theory tells us that this tradeoff comes with a net advantage.

It is a truism that for a fixed average power, the greater the bandwidth $W$ the greater the data rate $R$ that can be reliably achieved.\(^\text{14}\) This is true even in the presence of background radiation and noise, the explanation being that larger $W$ offers more degrees of freedom for representing data, overcoming any increase in noise due to the larger $W$. This is true as well at radio wavelengths (in the presence of white background noise) (Messerschmitt 2015). Larger $W$ is especially beneficial when, as here, background radiation is neglected, so that increasing $W$ does not increase the accompanying background rate.\(^\text{15}\)

The number of degrees of freedom for choosing codewords is $T_c W$, and the average detected photons per codeword is $T_s A_A$. Thus, the average detected photons per degree of freedom is $\Lambda_A W$, which decreases with $W$. Thus the available photons become more sparse over the available degrees of freedom as $K_s$ decreases and $W$ increases.

\(^\text{14}\) Of course increasing $W$ could not decrease the available $R$, since there is always the option to not make use of the extra bandwidth.

\(^\text{15}\) Of course the background will eventually become significant as $W$ is increased. Even in the regime of significant background the benefit of larger $W$ remains.
3.8. Technological limits on BPP

If a PPM modulation code is chosen, from (9) the fundamental limit on BPP increases monotonically as \( K_s \) decreases. From the relation \( \Lambda_P T_s = K_s \), since detected peak power \( \Lambda_P \) is determined by the fixed transmit peak power \( P_T \) together with the various sources of loss (propagation, atmosphere, quantum efficiency, etc.), it follows that our handle for reducing \( K_s \) is to reduce slot duration \( T_s \). The smallest feasible value of \( K_s \) is therefore determined by the smallest \( T_s \) that it is practical to implement.

This trade-off is also influenced by the data rate \( R \) and photon efficiency BPP we are attempting to achieve. For example, for \( R = 1 \) bps, \( PAR = 1000 \), and \( BPP = 10 \), then the average and peak detected powers must be \( \Lambda_A = 0.1 \) and \( \Lambda_P = 100 \) photons per s. If the smallest feasible value is \( T_s = 1 \) ns based on technological limits and power consumption considerations in probe electronics, then the result is \( K_s = 10^{-7} \) photons in one time slot on average. From (9) this would get us within one part in \( 5 \cdot 10^{-8} \) of the fundamental limit. This is way beyond a point of diminishing returns, and would create serious issues in the acquisition of the signal and the design of the ECC.

At the low data rates expected in this application, technological limitations on the generation of short pulses are not concerning. The capabilities of the ECC layer, subject to limitations in the available probe processing resources (and the electrical power for that processing) are expected to be limiting.

4. ERROR-CORRECTION CODING

The ECC coding layer in the transmitter (Figure 3) adds controlled redundancy and passes the resulting data to the modulation coding. The data rate of the coded bits is higher than the scientific data rate \( R \) due to the added redundancy. The ECC decoding layer in the receiver accepts an unreliable replica of that data from the modulation decoding, and, exploiting the included redundancy reconstructs a dramatically more reliable replica of the scientific data.

Structurally the modulation code layer deals with blocks of \( m \) bits. (If it is based on a PPM modulation code then \( m = \log_2 M \) where \( M \) is the number of timeslots per codeword.) The ECC layer constructs longer codewords by grouping \( L \) of these together, for a total of \( Lm \) bits per ECC codeword. Choosing larger \( L \) makes it possible to achieve higher reliability (all other parameters equal) if the codebook is well constructed.

The design of effective ECC codebooks for large \( L \) is a non-trivial task due to the exponential growth in the codebook size as \( L \) increases. In particular basing a codebook on storage and retrieval of codewords becomes impractical, and instead coding and decoding must be based on algorithmic processing. Those algorithms in turn are typically based on mathematical theory. That processing can be quite substantial, but fortunately by far the most intensive processing is on the decoding side, where it has no impact on the probe mass and energy supply.

Any chosen ECC scheme leaves us with a residual (non-zero) error probability. The goal in ECC design is, for a stated reliability objective, to accommodate scientific data at a rate approaching the fundamental limit of \( BPP = \log_2 M \) for an \( M \)-ary PPM modulation coding layer. In practice this requires that the ECC layer be split into sub-layers; that is, coding/decoding using different schemes (with different strengths) are applied sequentially. For example, if there are two such sub-layers, they are called the inner and outer codes. The objective of this section is to illustrate ECC rather than perform a full design of an optimum ECC scheme. Thus, we assume a single ECC coding-decoding layer, and demonstrate how it considerably improves reliability, but concede that it does not closely approach the fundamental limit.

4.1. An image example without ECC

A primary payload for early interstellar probes is the communication of images (Section 2.1). A reasonable reliability objective might be “a 99% probability that any transmitted image is received and decoded correctly”. A single image might contain 1 Mb of data (a million pixels at one bit per pixel after compression). Consider a PPM modulation code communicating \( m = 10 \) bits per PPM frame, and examine its performance in the absence of ECC. A 1 Mb image requires 100,000 PPM frames.

Recall that PPM without additional coding and with the neglect of background radiation suffers erasure errors. To decode an entire image correctly requires zero erasures in 100,000 PPM frames. We achieve this goal for 99% of images if

\[
(1 - \Pr\{\text{erasure}\})^{100,000} = 0.99,
\]

which can be satisfied by \( \Pr\{\text{erasure}\} = 10^{-7} \). We saw in Section 3.5.1 that for PPM without ECC this requires \( K_s \approx 16 \). The resulting photon efficiency is 10 bits per 16 photons, or \( BPP = 0.625 \), a substantial loss compared to the fundamental limit of \( BPP = 10 \) for this modulation coding scheme with \( PAR = 2^{10} \).

4.2. A Reed-Solomon ECC

To illustrate how high reliability can be achieved with considerably larger photon efficiency BPP, we will add a Reed-Solomon (RS) ECC. Although RS is not currently the state of the art, it is relatively simple and its performance is readily modeled. RS is also widely used, especially in older systems like the NASA Viking spacecraft and the DVD/Blueray video disks, and has relatively low-complexity encoder and decoder implementations that make it attractive for optical communications (Reed & Solomon 1960; Geisel 1990). RS is known for its powerful erasure correcting capability, which makes
it an attractive match to PPM modulation coding in the low background regime.

4.2.1. Structure

Consider an RS\{L, K, m\} code, which has three parameters \{L, K, m\}. Each codeword is constructed of a group of \(m\)-bit symbols, which map naturally onto \(M\)-ary PPM frames where \(M = 2^m\). Each RS codeword contains \(L\) such symbols, with the constraint \(L = M - 1\). Of the \(L\) symbols, \(K\) are drawn from the scientific data and \(L - K\) are redundant, meaning that they are calculated by applying a specified algorithm to the \(K\) scientific data symbols. By varying the value of \(K\), the amount of redundancy and hence the reliability of reconstructed scientific data can be varied. RS coding and decoding is based on the mathematics of Galois fields, which is the source of the constraint \(L = M - 1\).

With the background-free assumption, the only type of error is erasures, which have the special property that the decoder knows which symbols have been erased. RS codes are typically decoded algebraically by bounded minimum distance (BMD) decoders (Geisel 1990), which guarantee correct decoding whenever the number of erasures is \(L - K\) or fewer (assuming there are no other types of errors). Thus as we increase \(K\) the redundancy decreases and the erasure-correction capability also decreases.

4.2.2. Example

The RS(1023,\(K\),10) code is a natural match to a modulation-layer based on \(M\)-ary PPM with \(m = 10\). The total number of bits in each codeword is \(1023 \cdot 10 = 10230\), of which \(10 \cdot \(K\)\) are scientific data bits. For a given choice of \(K\) we can calculate the maximum RS decoding error rate that still allows the overall reliability objective (99% of images received correctly) to be achieved. From this we can establish the maximum erasure probability at the decoder input, and the corresponding number of photons per PPM timeslot \(K\), and the photon efficiency BPP.

Consider \(K = 255\), which can correct \(1023 - 255 = 768\) or fewer erasures. For this choice, each codeword contains \(10 \cdot 255 = 2550\) information bits and 7680 redundant bits. The number of RS codewords required to convey a 1 Mb image is \(10^6/2550 \approx 392\). To correctly decode the entire image with 99% probability requires

\[
(1 - \text{Pr\{RS decode error\}})^{392} = 0.99
\]

which is satisfied by

\[
\text{Pr\{RS decode error\}} = 2.56 \times 10^{-5}.
\]

For our illustrative RS(1023,255,10) code, it can be shown that this value of \(\text{Pr\{RS decode error\}}\) will be achieved when \(P_E = 0.694\) where \(P_E\) is the erasure probability. That is, the probability that there occurs more than 768 erasures within a codeword is about 2.56 \times 10^{-5} when \(P_E = 0.694\). From \(P_E = e^{-K_s}\), the required photons per PPM timeslot is \(K_s = 0.365\). This is considerably smaller than the \(K_s = 16\) needed without ECC.

The number of information bits conveyed per \(2^{10}\)-ary PPM frame is no longer 10 bits as in the case of PPM without ECC. Rather, in each RS codeword there are 255 10-bit PPM frames that represent the scientific data, out of a total of 1023 10-bit PPM frames. Thus the overall number of information bits per PPM frame is \(10 \cdot 255/1023 = 2.49\) bits. In other words, we require an increase in the number of PPM frames required to represent each image by a factor of \(10/2.49 = 4.01\). The resulting photon efficiency is\(\text{BPP} = 2.49/0.365 = 6.82\). This is a considerable improvement over PPM without ECC, but still falls somewhat short of the fundamental limit of\(\text{BPP} = 10\). This gap could be shrunk in a number of ways, including concatenating the RS code with an appropriate outer code, as discussed in Section 4.3.

This RS(1023,255,10) code illustrates that achieving high BPP with high reliability requires averaging over a large number of photon detection events. Per RS codeword there are 2550 bits of scientific data decoded, based on stochastic photon detection events averaging \(1023 \cdot 0.365 = 373.4\) in number.

4.2.3. Maximizing BPP

The foregoing calculation can be repeated for every possible choice of \(K\), with the result shown in Figure 6. The abscissa is the code rate \(K/L\) (fraction of coded symbols that are information symbols), and the ordinate is the photon efficiency BPP. This demonstrates an optimal code rate of 0.22 for which BPP = 6.834 bits per photon. At this code rate and BPP, the probability of erasure is \(P_E = 0.724\). In words, 72.4% of the PPM frames are not visible to the receiver because they have zero detected photons.

Figure 6 shows that BPP does not increase indefinitely as the code rate (and hence \(K_s\)) is decreased, which might have been expected from (9). This is due to the fixed length \(Lm = 10230\) of the RS(1023,\(K\),10) codewords, whereas (9) implicitly assumes that \(Lm \to \infty\). If we allow \(m\) to increase so as to increase the length of the RS codewords, the achievable optimum BPP increases as expected, and the peak occurs at ever smaller code rates (Figure 7).
4.2.4. *ECC efficiency*

An efficiency measure is the ratio of the actual BPP to the maximum BPP = \(m\) for \(M\)-ary PPM. For each \(m\) considered in Figure 7 the performance parameters of the code are tabulated here:

| \(m\) | 9  | 10  | 11  | 12  | 13  | 14  |
|------|----|-----|-----|-----|-----|-----|
| Maximum BPP | 5.513 | 6.834 | 8.187 | 9.545 | 10.889 | 12.207 |
| Optimum Pr {erasure} | 0.646 | 0.724 | 0.787 | 0.832 | 0.870 | 0.900 |
| Optimum \(K_s\) | 0.437 | 0.323 | 0.240 | 0.184 | 0.139 | 0.105 |
| Optimum code rate | 0.268 | 0.221 | 0.179 | 0.146 | 0.117 | 0.092 |
| Maximum ECC efficiency | 61.3% | 68.3% | 74.4% | 79.5% | 83.8% | 87.2% |

As expected, the efficiency increases with \(m\) due to the increasing codeword length of \(L\) PPM symbols (where \(L = 2^m - 1\)).

In Section 3.7.2 the numerical results were based on \(K_s = 0.1\). Using an RS code, this could be achieved with \(m = 14\) and a code rate of \(K/L = 0.092\), suggesting an RS(16383,1507,14) code. This design choice would entail two implementation challenges, a large PAR = 16384 and a large symbol size \(m = 14\). The processing load (particularly on the receive side) increases exponentially with \(m\).

4.3. *Other ECC approaches*

We have shown that our baseline RS coding example actually performs quite well, which is a testament to its erasure-countering capabilities matched to the erasure nature of PPM errors when background radiation is neglected. However, it is well worth exploring alternative ECC designs, with the goals of (a) achieving closer to 100% ECC efficiency and (b) doing this with a more modest PAR, such as PAR = 10

An ECC approach displaying high performance with PPM is SCPPM, which is the concatenation of a convolutional outer code and an accumulate PPM (APPM) inner code, with soft decision inputs and iterative decoding (Cheng et al. 2006) (Barsoum et al. 2007). Significant improvements are claimed over RS coding, although the comparisons are typically made against a sub-optimal RS code choice and not on the background-free erasure channel as considered here. Therefore the relative advantage of this approach over our baseline RS scheme has not been established. This is also the case for a range of other approaches that have been suggested for the optical PPM channel, including parallel concatenated convolutional and PCM codes (PCPPM) (Peleg 2000) and low-density parity check codes with PPM (LDPC-PPM) (Barsoum et al. 2007). Most ECC approaches considered for a PPM modulation code to date are binary codes adapted from other applications, which do not take advantage of the non-binary symbol alphabet of PPM. Development of non-binary variants of SCPPM, PCPPM and LDP-PPM would be a fruitful research direction.

In all cases, there is a fundamental advantage to codebooks with long codewords. Traditionally the limiting factor in the practical length of codes has been decoder complexity, which typically increases exponentially with code length. However, the recently proposed Fountain codes (MacKay 2005) offer the possibility of very long codes because of a decoding complexity that increases only linearly with code length. We are investigating the concatenation of a non-binary inner code with a Fountain outer code, potentially also incorporating iterative decoding. The goal is to achieve a scheme approaching 90% ECC efficiency with 10-ary PPM, resulting in BPP \(\rightarrow 9\) on the background-free erasure channel with this modest PPM order and PAR.

4.4. *Outage mitigation with ECC*

Outages (Section 2.5) are manifested by erasures, and thus are amenable to RS coding as described earlier. However, they are statistically grouped together, and interleaving in the transmitter and de-interleaving in the receiver spreads those erasures out in time, ensuring that all parts of the scientific data stream are equally impacted. The ECC design should take the increased erasure probability into account based on a worst-case assumption as to outage probability.

Outage events vary in severity in depth and duration, and can be represented by statistical models whose parameters vary depending on the receiving site location, time of year, and time of day. Precise modeling of outage statistics must await selection of a receiver site. To gain some initial insight, outages can be modeled as OFF-ON events, where loss is either total or non-existent. In this case the PPM decoding will observe outages as erasures, indistinguishable from erasures due to photon statistics. If we maintain RS-coding symbol integrity in the interleaving process described in Section 2.5, after de-interleaving RS symbol erasures due to the two mechanisms occur pseudo-randomly.

Consider a single receiver for which the long-term average outage probability is \(P_O\), or there are outages 100 \cdot \(P_O\)% of the time on average. A receiver site that experiences infrequent cloud-cover might have \(P_O < 0.1\), while a cloudier site might have \(P_O > 0.5\). Nighttime-

---

16 More sophisticated algorithms (e.g. (Schmidt et al. 2010)) are capable of correcting some fraction of erasure patterns that contain greater than \(L - K\) erasures. However, throughout this work we conservatively assume an erasure correcting limit of \(L - K\).
only operation could bring $P_O$ well above 0.5. The overall erasure probability at the PPM decoder output (RS decoder input) is $P_O + (1 - P_O) \cdot P_E$, where $P_E$ is the erasure probability during non-outage periods. Regardless of $P_E$, the overall erasure probability is never less than $P_O$.

The use of BPP as a measure of decoder performance (as in Figure 6) is difficult to interpret for outages. For this case we adopt an alternative approach. The same scientific data rate and probability of error following RS decoding can be maintained when outages are present if the receive-aperture area $A_R$ (and hence received power $\Lambda A$) is appropriately increased. In conjunction with an increase in $A_R$, the optimum level of redundancy in the RS coding must be increased (the code rate $K/L$ decreased) to accommodate the higher overall erasure probability. For an equivalent data rate, this results in a higher PPM frame rate, which in turn depends on the increase in receive power to achieve the same RS decoding error rate objective.

The required increase in $A_R$ relative to the zero outage case is illustrated in Figure 8 for different outage probabilities as a function of RS coding redundancy. Note that the optimum code rate decreases with increasing outage probability, consistent with the increased erasure probability due to the outages. After the most advantageous choice of code rate, the required increase in $A_R$ is roughly approximated by $1/(1 - P_O)$. For example, for $P_O = 0.5$, $A_R$ must be increased by a factor of approximately 2.3, or the aperture diameter must be increased by a factor of $\sqrt{A_R} \approx 1.5$.

The details on different types of background radiation are found in (Lubin et al. 2018). Our concern here is with quantifying the conditions under which such background can be neglected, as this is a basic assumption in our discussion of the modulation and ECC coding layer. The theory tells us that background radiation can be neglected (in the sense that it does not appreciably affect the fundamental limit of channel capacity) if $SBR = \Lambda_A/\Lambda_B \geq 10^2$ (Appendix C). Thus the size of the background power in relation to received probe-originating signal power is what matters.

5. BACKGROUND RADIATION

5.1. Optical bandwidth

Within the optical bandwidth $B$ of the signal presented to the optical detector, the power spectrum
of background radiation is essentially white, and thus $\Lambda_B \propto B$. Any expectation of negligible background is based on a small $B$. The relevant value of $B$ depends on the multiplexing method (Section 2.6). For $J_p$ downlink channels, the optical bandwidth for TDM or CDM will generally be $J_p$ times as large as with WDM. This can be misleading, however. With TDM the transmit power will be more than $J_p$ as large, restoring SBR, and with CDM the codeword correlation operation in the receiver should help to restore SBR. In the following we focus on the simplest case of WDM.

5.1.1. Signal bandwidth

The goal in bandpass filter design is to achieve $B \approx W$ for a signal bandwidth $W$. Adopting a PPM modulation code, this bandwidth is $W \approx \Delta P/K_s$ (Section 3.7.5). Since $P_p$ is assumed constant, and if $P_A$ is also constant, both $W$ and $\Lambda_A$ decrease as $d^2$. In that case SBR remains constant with distance $d$. However in practice optical filter bandwidth may not be able to track $W$ due to technological limitations and the lower bound of laser linewidth. As well, $P_A$ will decline due to RTG half-life. Thus, in practice we expect some deterioration in SBR during a decades-long communication phase of the mission.

5.2. Sources of background

Astronomical sources of background radiation divide into (a) resolved and (b) unresolved radiation. The intensity of all sources depends on wavelength, and estimates are listed in Table 6 for two wavelengths $\lambda_0 = 400 \text{nm}$ and $1.0 \mu$m. The largest (and thus most important) of these are, respectively, (a) the target star light and (b) the combination of Zodiacal light originating in the Earth and target solar systems and faint-star light originating from other more distant stars in our galaxy (Lubin et al. 2018). Coincidently, the latter two unresolved sources have essentially the same intensity at the wavelengths listed (although this does not hold true at other wavelengths). Within a narrow optical bandwidth $B$ all sources can be considered white, so that the total background radiation is proportional to $B$.

5.2.1. Target-star light

The target star has a large emission in total. The worst case assumption is made in Table 6 that the star falls entirely within the FOV of the receive aperture. The design of the receive aperture may attenuate this radiation, although this is a challenging design issue (Section 2.10). For wavelengths shorter than $1 \mu$m this radiation falls dramatically, and thus it is advantageous to choose $\lambda_0 = 400 \text{nm}$, which falls near the atmospheric cutoff of 330 nm.

The star appears to be a point source, with the totality of star radiation contributing to the background. The star radiation is isotropic, so that portion intercepted by the receive aperture is proportional to area $A_R$, and thus $\Lambda_B = \eta \Gamma_{B,s} A_R B$ where $\Gamma_{B,s}$ is a $\lambda_0$-dependent constant. This value does not depend on aperture synthesis, although synthesis can be helpful in rejecting some or all of this radiation. For Proxima Centari (a cool red dwarf) $\Gamma_{B,s}$ is several orders of magnitude smaller at the shorter wavelength. Thus the shorter wavelength is advantageous unless the aperture can reject much of this source of background.

Based on (1) $\text{SBR} \propto P_A A_f/B$ independent of $\eta A_R$, since the latter affects probe signal and star light equally. Thus, SBR can be increased by using larger transmit power $P_A$, larger transmit aperture area $A_f$, or reducing bandwidth $B$. For this source alone at $\lambda_0 = 400 \text{nm}$, the threshold of negligible background SBR $\approx 10^8$ can be achieved with $P_A = 13 \text{W}$ (a factor of 1300 larger than assumed earlier) for $B = 10 \text{kHz}$. At $\lambda_0 = 1 \mu$m this increases to an unreasonable $P_A = 8.1 \text{kW}$.

The general conclusion is that aperture partial rejection of target star emission (Section 2.10) is essential. Otherwise background radiation becomes quite significant, which falls beyond the scope of this paper.

5.2.2. Zodiacal and faint-star light

Zodiacal light, particularly with origins in our own Solar System, has approximately uniform luminosity across the entire FOV of the receive aperture. That is also true of Zodiacal light originating in the target solar system and faint-star light, although obviously there is some directivity for all those sources.

For each fully synthesized sub-array (assumed to be diffraction-limited) the total background power is independent of the synthesized sub-aperture area $A_R/N$. This is because increasing that area has two effects (larger gathering area and smaller FOV) that exactly offset. Thus, for Zodiacal or faint-star light the total background power is $\Lambda_B = N \eta \Gamma_{B,z} B$ where $\Gamma_{B,z}$ is a $\lambda_0$-dependent constant. The factor of $N$ results from the incoherent addition of background from the $N$ sub-apertures. As seen in Table 6 the value of $\Gamma_{B,z}$ is not strongly dependent on wavelength, but does decrease modestly at the shorter wavelength. Including both sources approximately doubles the total background power.

---

Table 6. Background radiation from two sources

| Source              | Units       | 400 nm      | 1.0 $\mu$m |
|---------------------|-------------|-------------|------------|
| Target-star light   | ph/s-m²-Hz  | $8.0 \cdot 10^{-10}$ | $2.0 \cdot 10^{-7}$ |
| Zodiacal light      | ph/s-Hz     | $1.1 \cdot 10^{-16}$ | $4.7 \cdot 10^{-15}$ |
| Faint-star light     | ph/s-Hz     | $1.1 \cdot 10^{-16}$ | $4.7 \cdot 10^{-15}$ |

---

17 Even shorter wavelengths would be advantageous for a space-based receive platform where atmospheric attenuation is not an issue.
The resulting SBR depends on the design parameters through
\[
\text{SBR} \propto \frac{P_A A_T}{B} \cdot \frac{A_R}{N}.
\]
Thus the SBR grows with larger synthesized sub-aperture area \(A_R/N\), since growing the diffraction-limited sub-apertures comes at no cost in terms of unresolved background radiation, but growing the total receive aperture area \(A_R\) by increasing \(N\) does come at a cost. For example, for \(D_R = 1\) km and \(N = 10^6\) the threshold of negligible background for this source alone occurs at \(P_A = 4.6 \mu \text{W}\), which is much smaller than assumed earlier. Thus design for a region wherein Zodiacal and faint-star light are negligible is relatively straightforward.

6. CONCLUSIONS

The design of a communication downlink from low-mass interstellar probes is extremely challenging. While the downlink appears to be feasible in theory, we have identified areas where technological innovation and invention will be needed, and the feasibility of those technologies remains in question. Thus this paper can serve as a roadmap to further investigation and research, and the ultimate outcome depends on the success of those efforts.

APPENDIX

A. DATA VOLUME VS. LATENCY

A.1. Scaling laws

Referring to Figure 1, since the total distance to the end of transmission is \(D_t + D_d\),
\[
T_t = (D_t + D_d) \left( \frac{1}{u} + \frac{1}{c} \right).
\]
(A1)

This latency does not depend on the data rate.

As governed by the square-law propagation loss, the total data volume is, when the initial data rate is \(\mathcal{R}\),
\[
V = \int_0^{D_d/u} \frac{\mathcal{R}}{(1 + ut/D_t)^2} dt = \frac{\mathcal{R}}{u} \cdot \frac{D_t D_d}{D_t + D_d}.
\]
(A2)

A direct relationship between \(T_t\) and \(V\) can be obtained by combining (A1) and (A2) while eliminating the variable \(D_d\). This can be expressed in terms of \(\{ \mathcal{R}_o, u_0, \xi \}\) by substituting the scaling relations given in Section 2.4.1. The maximum possible data volume is
\[
V \rightarrow \xi^{9/4} \cdot \frac{\mathcal{R}_o D_t}{u_0} \quad \text{as} \quad T_t \rightarrow \infty.
\]

A.2. Refinements

The previous results do not take into account the consequences of declining RTG electrical power during the transmission period. This is considered for the \(\xi = 1\) case, with results summarized in Table 7. Case A is the prediction of (A1) and (A2), which takes increasing distance and propagation loss into account.

We gratefully acknowledge funding from NASA NIAC NNX15AL.91G and NASA NIAC NNX16AL32G for the NASA Starlight program and the NASA California Space Grant NASA NNX10AT93H as well as a generous gift from the Emmett and Gladys W. Technology Fund in support of this research. PML gratefully acknowledges support from the Breakthrough Foundation for the Starshot program.

A web based calculator and more details on the NASA Starlight program can be found at: www.deepspace.ucsb.edu/projects/starlight.
B. SQUARE-LAW DETECTION

A real-valued passband signal \( x(t) \) centered at carrier frequency \( \nu = c/\lambda \) Hz is conveniently written in the form \( x(t) = \Re \{ s(t)e^{j2\pi\nu t} \} \) where \( s(t) \) is a complex-valued baseband signal, \( j = \sqrt{-1} \), and \( \Re \) denotes real part. Now consider two such signals \( x_1(t) \) and \( x_2(t) \) with carrier frequencies \( \nu_1 \) and \( \nu_2 \), both presumed to be much larger than the bandwidth of baseband signals \( s_1(t) \) and \( s_2(t) \). If these signals are added at an optical detector input, the output current is proportional to instantaneous input power,

\[
2 \cdot (x_1(t) + x_2(t))^2 \simeq |s_1(t)|^2 + |s_2(t)|^2 + 2\Re \{ s_1(t)s_2(t)e^{j2\pi(\nu_1-\nu_2)t} \},
\]

(B4)

where \( \pi \) denotes the complex conjugate of \( z \). The relation ‘\( \simeq \)’ indicates that out-of-band components with carrier frequencies \( 2\nu_1, 2\nu_2 \), and \( (\nu_1 + \nu_2) \) have been discarded.

B.1. Heterodyne

Assume that a large local oscillator signal \( x_1(t) = \sqrt{P_{LO}}e^{j2\pi\nu_0t} \) and input signal \( x_2(t) = \sqrt{P(t)}e^{j2\pi\nu_0t} \) with instantaneous power \( P(t) \) are mixed at the input to the square-law detector. Since \( P(t) \ll P_{LO} \), the \( P(t) \) term at the output can be neglected, and from (B4)

\[
2 \cdot (x_1(t) + x_2(t))^2 \simeq P_{LO} + 2\sqrt{P_{LO}P(t)} \cos 2\pi(\nu_0 - \nu_0). \]

Thus the signal term \( \sqrt{P(t)} \) (desirably amplified by \( \sqrt{P_{LO}} \)) appears principally at intermediate frequency \( (\nu_0 - \nu_0) \), which can be arranged to fall in the microwave spectrum. The d.c. term \( P_{LO} \) is an extra source of broadband noise (extending even to microwave) due to the shot-noise character of the local oscillator. This extra noise limits the photon efficiency that can be achieved.

B.2. Aperture synthesis

Consider a simple two-element aperture separated by distance \( b \) with an extra geometric delay \( \tau_g = b\sin\theta/c \) introduced to one element’s output signal due to incident angle \( \theta \) measured relative to a perpendicular bisector to the baseline of length \( b \). For synthesis introduce an instrumental delay \( \tau_i \) and phase shift \( \Phi \) (relative to carrier frequency \( \nu \)) at the output of the second element. Both elements observe the same baseband signal \( s(t) \) and the same carrier frequency \( \nu_0 = \nu_1 = \nu_2 \). Then the two baseband signals at the optical detector input are

\[
s_1(t) = \sqrt{P(t - \tau_g)}e^{-j2\pi\nu_0\tau_g} \quad \text{and} \quad s_2(t) = \sqrt{P(t - \tau_i)}e^{j(\phi - 2\pi\nu_0\tau_i)},
\]

where \( P(t) \) is the non-negative real-valued instantaneous input power. Then (B4) becomes

\[
2 \cdot (x_1(t) + x_2(t))^2 \simeq P(t - \tau_g) + P(t - \tau_i) + 2\beta P(t - \tau_g)P(t - \tau_i) \beta = \cos(\phi + 2\pi\nu_0(\tau_g - \tau_i)).
\]

The \( |\beta| \leq 1 \) term (called the fringe) controls the interference: none, constructive, or destructive. The key to successful synthesis is to control \( \beta \) through the (presumably adaptive) control of \( \tau_i \) and \( \phi \).

B.2.1. Coded waveform

The waveform \( P(t) \) generated by the coding layer is constructed by a superposition of ON-OFF pulses with duration \( T_s \) (Section 3.3.3). After two-aperture synthesis with \( \Delta \tau = \tau_g - \tau_i \), the resulting power waveform has three levels due to overlap (or not) of the pulses. If a single pulse has amplitude unity, the resulting pulse energy after synthesis is \( (1 + \beta(1 - \Delta \tau)) \). Generally we want \( \beta \approx 1 \) to yield a doubling of pulse energy, but that desired effect is diminished by \( \Delta \tau \neq 0 \).

The broadening of the pulse due to \( \Delta \tau \neq 0 \) causes an ON-pulse to spill out of one timeslot, and this interferes with the proper decoding of PPM frames. (The effect is similar to unwanted background radiation, albeit in a time- and data-dependent fashion.) To minimize this effect we need \( \Delta \tau \ll T_s \). Since the waveform bandwidth is of order \( W \approx T_s^{-1} \), it follows that we need \( W\Delta \tau \ll 1 \).

B.2.2. Target star radiation

This source of radiation has constant power \( P(t) = P_0 \), so that power after synthesis is \( P_0(1 + \beta) \). Therefore we want \( \beta \approx -1 \) to reject this radiation. Physically this phase modification shifts the diffraction pattern from constructive to destructive interference.

Two-aperture synthesis has insufficient degrees of freedom to simultaneously constructively amplify signal and destructively attenuate starlight background. This can be accomplished in principle by introducing more aperture elements, often on non-collinear baselines. The question is how accurately (and adaptively) phase matching can be performed for multiple aperture elements over a relatively large range of baseline lengths and in the presence of atmospheric turbulence.

C. DIRECT-DETECTION CAPACITY

Once the physical layer is established, its input-output relationship can be modeled statistically. For any such statistical model, the channel capacity \( C \) (Section 3.2) specifies the maximum information rate (in bits per second) that can be achieved by channel coding to an arbitrary level of reliability (Gallager 2008; Cover & Thomas 1991; Messerschmitt 2008). Actually achieving reliable rates in this region requires sophisticated channel coding, as
Figure 9. For a continuous-time photon-counting channel, a log-log plot of BPP vs SBR with the different curves corresponding to different values of PAR. The curves are labeled by \( \log_2 \text{PAR} \), so for example 14 corresponds to \( \text{PAR} = 2^{14} = 16,384 \). The axis \( \log_{10} \text{BPP} = 0 \) corresponds to a photon efficiency of one bit per photon. The axis \( \log_{10} \text{SBR} = 0 \) corresponds to equal average signal and background power.

illustrated by concrete example of a PPM modulation combined with ECC (Section 3.7).

We give three examples of statistical models and associated channel capacity in this paper, without derivation. These are the direct-detection (or photon-counting) channel considered here, the channel formed by heterodyne detection (Sections 2.14 and B.1), and the discrete-time channel observed by the ECC layer with an underlying PPM modulation code (Section 3.7).

C.1. Capacity result

The direct-detection channel is modeled by photon detection events governed by Poisson arrival statistics. This is the most random arrival process, in which inter-arrival times are statistically independent and obey an exponential distribution. The capacity has been determined for this statistical model under the peak and average power constraints described in Section 3.2.2. Define two parameters

\[
s = \frac{1}{\text{PAR} \cdot \text{SBR}}, \quad q = \min \left\{ \frac{1}{\text{PAR}}, \frac{(1+s)^{1+s}}{e^{-s}} - s \right\},
\]

and then (Wyner 1988)

\[
\frac{C}{\Lambda_A} = \text{BPP} = \text{PAR} \cdot \log_2 \frac{(1+s)^{q(1+s)}s^{(1-q)s}}{(q+s)^q s^{(q+s)}}.
\] (C5)

In the limit as SBR → ∞ (or equivalently \( s \to 0 \)), (C5) simplifies to (yielding (5))

\[
\text{BPP} \to -\text{PAR} \cdot q \log_2 q
\] (C6)

\[
q \to \min \left\{ 1/\text{PAR}, 1/e \right\}.
\]

C.2. BPP is unbounded

Notably the capacity \( C \) is unbounded even for finite average power,

\[
\text{BPP} \to \infty \text{ as PAR} \to \infty \text{ for any SBR} < \infty.
\] (C7)

This confirms that a PAR constraint is necessary to prevent infinite BPP, and more importantly indicates that increasing \( \Lambda_p \) increases BPP and \( C \) even as \( \Lambda_A \) is held constant. (The uncertainty principle will intervene at very high PAR (Butman et al. 1982).)

C.3. Background-free regime

A log-log plot of BPP vs SBR is shown in Figure 9 over a wide range (12 orders of magnitude) of SBR.

This paper neglects the background radiation, which corresponds to \( \text{SBR} \to \infty \). In Figure 9 the value of BPP is nearly constant for \( \text{SBR} > 10^2 \), giving us a criterion for neglecting background. From (5), in this background-free regime \( \text{BPP} \approx \log_2 \text{PAR} \). Notably at large PAR there is reduced dependence of BPP on SBR in the regime where background radiation is significant.

REFERENCES

Barsoum, M., Moision, B., & Fitz, M. 2007, in Information Theory Workshop, 2007. ITW’07. IEEE (IEEE), 66–71

Burke, B. F., & Graham-Smith, F. 2009, An introduction to radio astronomy (Cambridge University Press)

Butman, S., Katz, J., & Lesh, J. 1982, IEEE Transactions on Communications, 30, 1262

Cash, W. 2011, The Astrophysical Journal, 738, 76

Cheng, M. K., Moision, B. E., Hamkins, J., & Nakashima, M. A. 2006, in Global Telecommunications Conference, 2006. GLOBECOM’06. IEEE (IEEE), 1–5

Cover, T., & Thomas, J. 1991, Elements of Information Theory (New York: John Wiley and Sons)

Dolinar, S., Birnbaum, K. M., Erkmen, B. I., & Moision, B. 2011, in Space Optical Systems and Applications (ICSOS), 2011 International Conference on (IEEE), 269–278

Dolinar, S., Moision, B., & Erkmen, B. 2012

Fountain, G., Kusnierkiewicz, D., Hersman, C., et al. 2009, The New Horizons spacecraft, New Horizons (Springer), 23–47

Gallager, R. G. 2008, Principles of digital communication (Cambridge, UK: Cambridge University Press)

Geisel, W. A. 1990
Giovannetti, V., Guha, S., Lloyd, S., et al. 2004, Physical Review Letters, 92, 027902
Gordon, J. P. 1962, Proceedings of the IRE, 50, 1898
Hemmati, H. 2006, Deep space optical communications, Vol. 11 (John Wiley & Sons)
—. 2009, Near-earth laser communications (CRC press)
Hippke, M. 2017, arXiv preprint arXiv:1706.03795
Lo, A. S., Glassman, T., Lillie, C., & McGraw, P. J. 2007, in Proc. Of SPIE, Vol. 6687, 668716
Lubin, P. 2016, JBIIS, 69, 40
Lubin, P., Brashears, T., & Knowles, P. 2018, The Path - Directed Energy for Interstellar and Rapid Interplanetary Missions
Ludwig, R., & Taylor, J. 2016, Voyager telecommunications (John Wiley and Sons, Inc)
MacKay, D. J. 2005, IEE Proceedings-Communications, 152, 1062
Messerschmitt, D. G. 2008, Some Digital Communication Fundamentals for Physicists and Others, Tech. Rep. UCB/EECS-2008-78, EECS Department, University of California, Berkeley
—. 2013, End-to-end interstellar communication system design for power efficiency, Tech. Rep. arxiv:1305.4684
—. 2015, Acta Astronautica, 107, 20
—. 2017, Proc. IEEE, 105, 1511
Peleg, M. 2000, Transactions on Emerging Telecommunications Technologies, 11, 475
Reed, I. S., & Solomon, G. 1960, Journal of the society for industrial and applied mathematics, 8, 300
Roh, W., Seol, J.-Y., Park, J., et al. 2014, IEEE communications magazine, 52, 106
Schmidt, G., Sidorenko, V. R., & Bossert, M. 2010, IEEE Transactions on Information Theory, 56, 5245
Smith, B. A., Soderblom, L. A., Bunfield, D., et al. 1989, Science (New York, N.Y.), 246, 1422, IR: 20100608, JID: 0404511; 1989/12/15 00:00 [pubmed]; 1989/12/15 00:01 [medline]; 1989/12/15 00:00 [entrez]; ppublish
Spencer, D. T., Bauters, J. F., Heck, M. J., & Bowers, J. E. 2014, Optica, 1, 153
Spencer, D. T., Tang, Y., Bauters, J. F., Heck, M. J., & Bowers, J. E. 2012, in Photonics Conference (IPC), 2012 IEEE (IEEE), 141–142
Traub, W. A., & Oppenheimer, B. R. 2010, Direct imaging of exoplanets, ed. S. Seager, Exoplanets (University of Arizona Press), 111–156
Wyner, A. D. 1988, IEEE Transactions on Information Theory, 34, 1449