Cascaded superconducting junction refrigerators: optimization and performance limits

A. Kemppinen, E. Ronzani, E. Mykkänen, J. Hätinen, J. S. Lehtinen, and M. Prunnila

VTT Technical Research Centre of Finland Ltd, 02150 Espoo, Finland

(Dated: August 6, 2021)

We demonstrate highly transparent silicon–vanadium and silicon–aluminum tunnel junctions with relatively low sub-gap leakage current and discuss how a tradeoff typically encountered between transparency and leakage affects their refrigeration performance. We theoretically investigate cascaded superconducting tunnel junction refrigerators with two or more refrigeration stages. In particular, we develop an approximate method that takes into account self-heating effects, but still allows to optimize the cascade a single stage at a time. We design a cascade consisting of energy-efficient refrigeration stages, which makes cooling of, e.g., quantum devices from above 1 K to below 100 mK a realistic experimental target.

Normal metal – insulator – superconductor (NIS) and semiconductor – semiconductor (Sm–S) tunnel junctions can be used for cryogenic electrical refrigeration, because the superconducting energy gap $\Delta$ and a bias voltage $V \lesssim \Delta/e$ allow thermionic energy filtering of the tunneling electrons [1–4]. Here $e$ is the elementary charge. Suspended lateral cold finger assemblies have been used to refrigerate electrically both electrons and phonons of macroscopic objects in proof-of-concept experiments [5–7]. Despite extensive efforts and potential for low-cost, compact, and maintenance-free devices, electrical refrigerators have not replaced conventional techniques such as dilution refrigeration. Challenges for practical applications include limited cooling power and complicated engineering of phonon and electron–phonon heat flows [8–11], and the limited temperature range for refrigeration that depends on $\Delta$. The latter could in principle be overcome by multi-stage coolers that utilize superconductors with different $\Delta$ [12–14].

Recently, the electronic refrigeration of a macroscopic silicon chip was demonstrated using Sm–S (Al–Si) junctions [15]. They were used as the cooling element and mechanical support, and as a blockade for phonon heat transport based on the Kapitza resistance [16] between Al and Si. This approach avoids complex arrangements of cold fingers, which should allow a simple multi-stage assembly, possibly even 3D integration, as Fig. 1(a) shows. Reference [15] concludes that refrigeration from above 1 K to below 100 mK is a realistic target, but requires significant further development: (i) Since both cooling power and phonon heat leaks are proportional to the tunnel junction area $A$, it is beneficial to decrease the characteristic resistance $R_A = R_T A$ to improve the ratio between cooling power and phonon heat leaks. Here, $R_T$ is the tunneling resistance of the junction. (ii) Suppression of phonon thermal conductance at temperatures above 500 mK requires other phonon blocking methods in addition to the Kapitza resistance, e.g., nanowire constrictions [17]. (iii) A superconductor with larger $\Delta$ than aluminum, e.g., vanadium, is needed for refrigeration above about 500 mK.

In this letter, we demonstrate both Si–Al and Si–V tunnel junctions with small $R_A$ and discuss the optimization of a cascaded cooler with multiple refrigeration stages. In particular, we evaluate how the thermal balance of the device depends on finite cooling efficiency, i.e., that cooling one side of a junction with power $P_C$ causes heating power $P_H$ on the other side. We use Fig. 1(a) as our model system, but our method is applicable also for more complex designs.

An important limitation for cooling with NIS or Sm–S tunnel junctions, especially at temperatures well below the critical temperature of the superconductor $T \ll T_c$, is the leakage resistance $R_0$ in the sub-gap regime. The sub-gap leakage can originate from nonidealities of the superconductor or the tunnel junction, but the fundamental limit is set by Andreev...
According to theory for opaque NIS junctions [21], the sub-gap leakage parameter \( \gamma = R_0 / R_T \) due to Andreev reflection is inversely proportional to the characteristic resistance: \( \gamma = A_{ab} R_K / (4 R_A) \). Here, \( A_{ab} \) is the effective tunneling area per conductance quantum, \( R_K = h / e^2 \) is the quantum resistance, and \( h \) is the Planck constant. Andreev reflection thus prevents having low \( \gamma \) and \( R_A \) simultaneously. The observed \( A_{ab} \) of Andreev limited Cu–AlO–Al junctions is between about 10...30 nm², factor 5...10 higher than theoretical estimates, which has been explained by inhomogeneities of the tunnel barrier [21][24]. There is no reference value or theoretical prediction for \( A_{ab} \) in Sm–S junctions, which possibly depends on the microscopic quality of tunnel junctions. Therefore we define a phenomenological leakage area,

\[
A_\gamma = \frac{4 \gamma R_K}{R_A},
\]

and use it as a figure of merit for tunnel junctions. Parameter \( A_\gamma \) yields an upper limit for \( A_{ab} \), and allows empiric means for comparing leakage strength in junctions with different \( R_A \) in view of the fundamental trade-off between \( \gamma \) and \( R_A \).

Figure 1(b) shows the simulated relative cooling \( T_N / T_S \), i.e., the ratio between temperatures of the Sm and S sides for a Si–V junction at \( T_S = 1 \) K and with superconducting gap is \( \Delta_V = 600 \mu eV \) [25] \((T_S \approx 4 \) K). The thermal power of phonon leakage that limits cooling is

\[
P_L = g_{ph} A(T_S^4 - T_N^4),
\]

which is valid for several processes, but the prefactor \( g_{ph} = 160 \) W/m²K⁴ has been measured for the Kapitza resistance of Al–Si interface [15][16]. Figures 1(c–d) show similar simulations at 1 K and 0.5 K, respectively, where \( g_{ph} = 16 \) W/m²K⁴, obtained, e.g., with phonon engineering techniques [15]. We illustrate the limiting tradeoff between \( \gamma \) and \( R_A \) for 3 example values of \( A_\gamma \). Lower \( R_A \) at cost of higher \( \gamma \) is preferred when \( P_L \) is high, i.e., for large \( g_{ph} \) and \( T_S \). Smaller \( g_{ph} \) or \( T_S \) shift the preference to higher \( R_A \) with lower \( \gamma \), but overall Figs. 1(b–d) motivate the fabrication of tunnel junctions with \( R_A \leq 100 \mu m² \).

Figure 2(a) shows the voltage–current characteristics of Si–Al and an Si–V junctions fabricated with a similar process as in Ref. [18], but with increased junction transparency. The Si–Al junctions have 500 nm thick Al as superconductor, whereas Si–V junctions have a multilayer stack Si – SiOₓ – Al (25 nm) – V (150 nm) – Al (400 nm). High-quality vanadium junctions require a thin layer of, e.g., Al, at the junction interface [12]. Changing the thicknesses of V and the interface layer allows tuning \( \Delta \) roughly between those of Al and V, \( \Delta_{Al} = 200 \) µeV and \( \Delta_V = 600 \) µeV, respectively. The thicker Al layer relieves the multilayer film strain and improves quasiparticle diffusion due to its lower \( \Delta \) and normal-state resistivity compared to V. The Si–Al junction has \( R_A \approx 48 \mu m² \), which lower by factor 10 than the results of Ref. [15]. Yet it still has a relatively low \( \gamma = 8 \times 10^{-3} \), yielding \( A_\gamma = 62 \mu m² \). The same parameters for the Si–V junction are \( R_A = 71 \mu m² \), \( \gamma = 6 \times 10^{-3} \), and \( A_\gamma = 65 \mu m² \). These are significant improvements to previous results with Sm–S junctions [15] [18]. Notably, typical experiments of NIS junctions have about an order of magnitude higher \( R_A \), which may explain correspondingly lower values of \( \gamma \) [20][21].

Literature on Sm–S or NIS coolers usually aims at minimizing \( T_N / T_S \) as in Fig. 1 but also energy efficiency is crucial for a practical refrigerator. In the cascade of Fig. 1(a), the heating power \( P_{H,V} \), caused by cooling stage \( n \), must not overload the subsequent stage \( n-1 \). Unfortunately, the numerical optimization of the cascade is computationally demanding due to the large number of parameters that affect its performance. Neither does that provide physical insight into the system. Therefore, we develop an approximate method for optimizing the cascade a single stage at a time.

The heat currents to the normal (N) and superconducting (S) sides of an Sm–S or NIS junction and for tunneling to (+) and from (−) the normal side are

\[
\begin{align*}
\dot{Q}_N^+ &= \pm \frac{\Delta^2}{e R_A} \int_{-\infty}^{\infty} d\epsilon \rho_s(\epsilon, \gamma) f_S(\pm \epsilon, \gamma) f_N(\pm \epsilon \mp \nu, \gamma) \gamma (\epsilon - \nu), \\
\dot{Q}_S^- &= \pm \frac{\Delta^2}{e R_A} \int_{-\infty}^{\infty} d\epsilon \rho_s(\epsilon, \gamma) f_S(\pm \epsilon, \gamma) f_N(\pm \epsilon \mp \nu, \gamma) \gamma .
\end{align*}
\]

Here, \( \epsilon = E / \Delta \) and \( \nu = eV / \Delta \) are normalized energy and voltage, respectively, \( f_N \) and \( f_S \) are Fermi functions at normalized temperatures \( \tau = T / \Delta \). We use the Dynes density of states \( n_S \) to describe \( R_0 \) [26]. Ideally, for any target temperature \( T_N \) we keep the integral parts of Eqs. (3) constant, in their optimal value, by using fixed \( \nu \) and by choosing an optimal superconductor with \( \Delta_{opt} \approx T_N, T_S \) [15] that fixes \( \tau \). This is possible, e.g., by utilizing the proximity effect. Note that Eqs. (3) are insensitive to \( T_S \) when \( T_S \ll T_N \), as then the quasiparticle density is small. Therefore, when \( \Delta \) is scaled with \( T_N \), both \( P_C = -\dot{Q}_N^+ - \dot{Q}_N^- \) and \( P_H = \dot{Q}_S^- + \dot{Q}_S^+ \) scale as \( \propto T_N^2 \). This suggests that if a refrigeration stage yields a tem-
temperature ratio $T_n^2/T_{n-1}^2 \leq P_C/P_H$, the subsequent stage of a cascade that provides temperature $T_N$ may have sufficient cooling power to mitigate $P_H$, see Fig. 1(a). If the refrigeration performance does not satisfy this condition, excessive heating can be compensated by scaling tunnel junction areas, i.e., by having $A_1 > A_2 > A_3$ \ldots because the cooling and heating power of stage $n$ are proportional to its area, i.e., $P_{C,n}, P_{H,n} \propto A_n$.

The maximum heat load tolerated by a refrigeration stage, $P_{in} = P_C - P_L$ also depends on $P_L$. Unfortunately, $P_L$ of Eq. (2) follows a $T^4$ power law and increases faster as function of temperature than $P_C$. Thus cooling becomes increasingly difficult at higher temperatures even if materials with suitable $\Delta$ are available. However, thermal balance adjusts $T_N$ so that $P_C \geq P_L$. Hence we expect that the proportionality $P_C \propto T_N^2$ will approximately determine the magnitude of cooling power available for external load so that $P_{in} \approx T_N^2$. Therefore, when stage $n$ is the dominant heat load to stage $n-1$, i.e., $P_{H,n} = P_{in,n-1}$, heating of stage $n$ can be compensated with the ratio of areas

$$a_n = A_{n-1}/A_n \approx P_{in,n-1}/P_{in,n}^2 = P_{H,n}/P_{in,n}^2 \equiv P_{H,n}/P_{in,n}^2 = a.$$ (4)

The right hand side is written as the function of parameters of a single stage, i.e., this requirement allows to consider the performance of a single stage at a time.

An optimal refrigerator reaches the target temperature $T_T$ from $T_0$ with minimum $P_{H,n}/F_{in,n}$, i.e., it maximizes the cooling power with respect to the heating power. Since cooling and heating powers are proportional to the tunnel junction areas, our goal is to minimize $\prod_{n=1}^N a_n$. We do not know a priori whether it is optimal to have many stages (large $F$) with minimal $a$ or, conversely, small $F$ with larger $a$. Therefore we find an approximate optimum by considering $F$ stages that are equal on the logarithmic scale, i.e., have the same $a$ and $T_n/T_N$. Then $T_0/T_T = (T_n/T_N)^F$ and $F = \lg(T_n/T_T)/\lg(T_n/T_N)$, which yields $\prod_{n=1}^N a_n = a^F = O(\lg(T_n/T_N)^{-1})$, where $O$ is an optimization parameter we aim to minimize:

$$O \equiv a^{-\frac{1}{\lg(T_n/T_N)}} = \left(\frac{P_{H,n}^2}{P_{in,n}^2T_N^2}\right)^{\frac{1}{\lg(T_n/T_N)}}.$$ (5)

Note that $O$ summarizes the performance of each individual stage based on physical quantities (power and temperature) that characterize it. Therefore $O$ allows to optimize refrigeration stages independently of others. There is no direct dependence on phenomenological modeling parameters such as $\gamma$ or $g_{ph}$, and a low $O$ may be obtained in several ways. Parameter $O$ is a more comprehensive figure of merit compared to the typical optimization of $T_n/T_S$ shown in Figs. 1b–d) since it also takes into account the energy efficiency. To gain intuitive insight to the magnitude of $O$, let us consider a refrigerator with $T_1 = 1 \text{ K}$ and $T_F = 100 \text{ mK}$. If stages $n = 2, \ldots, F$ have the same $O = O_n$, the cascade then requires $A_1/A_F = O^{\lg(1 \text{ K}/100 \text{ mK})} = O$.

If the contribution of $P_L$ is small, for example because of good phonon filtering or low $T_S$, then $P_{H,n}$ dominates the heat balance of stage $n-1$, i.e., $P_{H,n} \approx P_{C,n-1}$, which yields the scaling $\gamma^2 = P_H T_n^2/(P_C T_S^2)$ instead of Eq. (4) and $O^\gamma = (\gamma^2)^{1/(\lg(T_n/T_S))}$ instead of Eq. (5).

Next, we study by simulations, how $v, \gamma, R_A, g_{ph}, \Delta, T_S$, and $T_N$ affect $O$ for Si–Al and Si–V junctions. We use Eq. (1) and a targeted value $A_T = 15 \text{ nm}^2$ to calculate $R_A(\gamma)$, which eliminates one dimension of the optimization problem. We focus on 3 specific values of $\Delta$, motivated by our experiments: 200, 450, and 600 $\mu$eV. This approach requires computing $P_C, P_H$, and $P_L$ in 5-dimensional space ($v, \gamma, g_{ph}, T_S, T_N$) for each $\Delta$ separately, which is computationally demanding, but dramatically easier than varying all parameters of the full cascade. More dimensions can be taken into account when optimising $O$, but here we focus on what we consider as the simplest realistic model. For each set of $v, \gamma, g_{ph}$, and $T_S$ we first find the balance temperature $T_{N,b}$ where $P_C = P_L$, and then temperature $T_{N,O}$ in range $T_{N,b} \leq T_N \leq T_S$ that yields minimum $O$ (alternatively $O^\gamma$). This computation allows us to determine, e.g., the requirements of efficient refrigeration for each value of $\Delta$. Finally, we discuss the optimization of a specific 3-stage refrigerator and compare our approximate results to simulations of the complete thermal balance.

Figures 3 (a–c) show $T_N/T_S$, $O$, and $O^\gamma$, respectively, as a function of $\gamma$ and $v$ for Si–Al stage at 0.3 K. The minimum of $O$ and $O^\gamma$ are obtained at smaller $v$ than the minimum of $T_N/T_S$, which shows that it is beneficial to sacrifice some of
the cooling power to improve the cooling efficiency. Furthermore, $O^1$ yields a smaller optimum $\gamma$, which is due to the fact that it emphasizes the internal efficiency of the stage, i.e., $P_H$ required to yield a low $T_N$ whereas $O$ optimizes the performance against $P_R$.

Figures 3(d–f) show $O$ for Al-based ($\Delta = 200 \, \text{µeV}$) and V-based ($\Delta = 450, 600 \, \text{µeV}$) stages as a function of $T_S$ and $g_{\text{ph}}$. Here $\gamma$ and $\nu$ have been optimized for each combination of $T_S$ and $g_{\text{ph}}$ as in Fig. 3(b). Regimes with $O \lesssim 10^2$ indicate roughly where Al- and V-based stages yield reasonable performance for practical applications. These results demonstrate how an increased $\Delta$ allows cooling at higher temperatures, but also how phonon heat conductance weakens the performance at higher temperatures. Aluminum can be efficient for $T_S \lesssim 500 \, \text{mK}$, which indicates that vanadium with 3 times higher $\Delta$ is applicable up to $T_S \lesssim 1.5 \, \text{K}$, but only if $g_{\text{ph}}$ can be made sufficiently low.

Figure 4 shows outcomes from the optimization along the cross-sections indicated with lines in Figs. 3(d–f). Figure 4(a) shows that factor 4 improvement of $g_{\text{ph}}$ can suppress $O$ by factor 100 for a V-based stage, but the difference is smaller for an Al-based stage. This means that the high $P_R$ of high temperature can make refrigeration extremely inefficient. Figure 4(b) shows that $a$ is less sensitive than $O$. Inefficient cooling thus yields a small temperature change per stage and increases $F$ rather than $a$. Figure 4(c) shows that high $T_S$ and $g_{\text{ph}}$ yield high optimal $\gamma$, which is in agreement with Fig. 4(a) and expected since low $R_1$ maximizes $P_C$. Figure 4(d) indicates that cooling from $T_0 = 1.2 \, \text{K}$ down to $T_1 = 100 \, \text{mK}$, i.e., from $T_0$ of a $^4\text{He}$ refrigerator is a realistic target for a 3-stage refrigerator if $g_{\text{ph}}$ can be suppressed below that of Kapitza resistance.

Finally, we evaluate the 3-stage refrigerator by simulating its complete thermal balance that depends on 15 fixed parameters. Table I shows a comparison of refrigerators A and B, based on minimizing $O$ and $T_N/T_S$ respectively. The parameters of refrigerator B were obtained similarly as explained above, but using $T_N/T_S$ is the optimization criterion instead of $O$. For both A and B, the parameters were obtained iteratively: We first optimized stages 2 and 3 for starting temperatures $T_1$ and $T_2$ obtained as in Fig. 4(d). Then we obtained $T_1 \ldots T_3$ from the thermal balance and used them to re-optimize stages 2 and 3. Such iteration is not necessary for refrigerator A, since the $O$-based estimate of Fig. 4(d) predicts temperatures quite accurately. We used Eq. 4 to determine $a_0$ also for refrigerator B since non-optimal scaling can yield dramatically worse performance. Note that when the inefficiency of refrigerator B is mitigated with larger $A_n$, it yields almost the same $T_3$ as refrigerator A. In these simulations, we set a reasonable heat load $P_{\text{in,3}}/A_3 = 1 \, \mu\text{W/mm}^2$, which results in the output power $P_{H,1}$ of 50 mW and 240 mW per mm$^2$ of $A_3$ for refrigerators A and B, respectively. This indicates that our method can suppress the energy consumption of the 3-stage refrigerator by about 80% compared to conventional optimization. For a small $P_{\text{in,3}}$, even lower $P_H,1$ can be obtained if the last stage is optimized using $O^1$.

To conclude, we have developed an optimization method for cascaded electronic refrigerators based on NIS or Sm–S tunnel junctions. This method helps to design energy-efficient refrigeration stages that mitigate heating generated in the device. Our results indicate that a practical electrical refrigerator that can cool, e.g., quantum devices from above 1 K to 100 mK is a realistic target, but requires phonon engineering and high-quality tunnel junctions. For the latter goal and for enabling cascaded junction refrigerators we have demonstrated highly transparent Si–Al and Si–V tunnel junctions with relatively low $\gamma$.

**ACKNOWLEDGMENTS**

This research was funded by the European Union’s Horizon 2020 research and innovation programme under grant agreements Nos 766853 EFINED (http://www.efined-h2020.eu/), Academy of Finland through projects ETHEC, UQS, and QuMOS (Nos 322580, 310909 and 288907) and Centre of Excellence program Nos 336817 and 312294.

---

**TABLE I. Parameters and performance of example refrigerators A and B starting from $T_0 = 1.2 \, \text{K}$ with $\gamma$ and $\nu$ based on minimizing $O$ and $T_N/T_S$, respectively, for each stage separately.**

| $n$ | $g_{\text{ph}}$ (W/m$^2$K$^4$) | $A_n/A_3$ | $\gamma$ ($10^{-3}$) | $\nu$ | $T$ (mK) |
|-----|-----------------|------------|-----------------|-------|---------|
| 1   | 40              | 32         | 5               | 0.88  | 598     |
| 2   | 40              | 8          | 17              | 1.6   | 196     |
| 3   | 160             | 1          | 1               | 0.8   | 89      |
| 4   | 160             | 1          | 1               | 0.8   | 91      |
DATA AVAILABILITY

Data available on reasonable requests from the authors.