Scaling Up Large-Scale Graph Processing for GPU-Accelerated Heterogeneous Systems

Xianliang Li
Huazhong University of Science and Technology
xianliang@hust.edu.cn

Abstract—Not only with the large host memory for supporting large-scale graph processing, GPU-accelerated heterogeneous architecture can also provide a great potential for high-performance computing. However, few existing heterogeneous systems can exploit both hardware advantages to enable the scale-up performance for graph processing due to the limited CPU-GPU transmission efficiency.

In this paper, we investigate the transmission inefficiency problem of heterogeneous graph systems. Our key insight is that the transmission efficiency for heterogeneous graph processing can be greatly improved by simply iterating each subgraph multiple times (rather than only once in prior work) in the GPU, further enabling to obtain the improvable efficiency of heterogeneous graph systems by enhancing GPU processing capability. We therefore present Seraph, with the highlights of pipelined subgraph iterations and predictive vertex updating, to cooperatively maximize the effective computations of GPU on graph processing. Our evaluation on a wide variety of large graph datasets shows that Seraph outperforms state-of-the-art graph processing. Our evaluation on a wide variety of large graph datasets shows that Seraph outperforms state-of-the-art graph processing. Our evaluation on a wide variety of large graph datasets shows that Seraph outperforms state-of-the-art graph processing. Our evaluation on a wide variety of large graph datasets shows that Seraph outperforms state-of-the-art graph processing. Our evaluation on a wide variety of large graph datasets shows that Seraph outperforms state-of-the-art graph processing. Our evaluation on a wide variety of large graph datasets shows that Seraph outperforms state-of-the-art graph processing. Our evaluation on a wide variety of large graph datasets shows that Seraph outperforms state-of-the-art graph processing. Our evaluation on a wide variety of large graph datasets shows that Seraph outperforms state-of-the-art graph processing. Our evaluation on a wide variety of large graph datasets shows that Seraph outperforms state-of-the-art graph processing. Our evaluation on a wide variety of large graph datasets shows that Seraph outperforms state-of-the-art graph processing. Our evaluation on a wide variety of large graph datasets shows that Seraph outperforms state-of-the-art graph processing. Our evaluation on a wide variety of large graph datasets shows that Seraph outperforms state-of-the-art graph processing.

I. INTRODUCTION

Graph is a well-known data structure that can represent a wealth of relationship between objects. Graph processing has a great potential to solve many real-world problems, e.g., path navigation, social network analysis, and financial fraud detection. As the graph size is increasingly growing up, it has become a critical turning point regarding how to store as well as further process these large-scale graphs efficiently.

A typical solution for large-scale graph processing is to divide the entire graph into many subgraphs that are then distributed onto different machines for the computation [1], [2]. Though these distributed systems (with more computation resources and storage resources) have made the impressive progress [3], [4], people are often more inclined to use a single-machine processing system, which is easier to manage and understand [5].

A wide spectrum of graph systems have emerged for processing large-scale graphs under a single machine [6], [7], particularly in the aspect of GPU acceleration because of its powerful computing capacity [8]–[10]. With an elegant advance-filter programming model, Gunrock [8] naturally integrates many well-optimized graphics analysis techniques. However, neither of these graph systems are able to handle the large-scale graphs that can not fit into the GPU global memory.

In an effort to cope with this problem, researchers extend to store the graph data in the large host memory for assisting the GPU computing. With the vertex data stored in the GPU memory, GTS [11] streams the subgraph data in an asynchronous manner. GraphReduce [12] only transfers the subgraph that has at least one active vertex or edge to the GPU. Unfortunately, due to relatively-low interconnect bandwidth between host and GPU (e.g., ~12GB/s for PCI-Express 3.0), the potential limitation of GPU accelerator under the heterogeneous architecture can become more serious.

Our motivating study (discussed in Section 2.2) also shows that, 75% of GPU computing capability can be underutilized even in the presence of existing state-of-the-art GPU-specific optimizations. It gradually becomes of great importance and necessity to scale up the performance of heterogeneous graph systems for large-scale graph processing. In this paper, we focus on studying whether and how we can provide a scale-up efficiency of heterogeneous graph systems under a commodity heterogeneous architecture.

Recently, there still exist a number of graph systems that attempt to improve the performance of heterogeneous graph systems. Graphie [13] proposes two renaming algorithms to improve the memory access efficiency, and keeps track of the active partitions to avoid moving the inactive partitions to GPU. Garaph [14] reduces the transmission amount by performing a part of sparse vertex updating on the host side. Nevertheless, for many real-world large graphs that can be dense and always active, these recent advances may still involve non-trivial amount of data transmission, making them limited for practical use.

In this paper, we present Seraph, a novel heterogeneous graph system, which can significantly scale up the performance of out-of-GPU-memory graph processing. The key insight of this work lies in a fact that each subgraph in one transmission iteration for many graph algorithms (e.g., BFS and SSSP) may involve much information that is useful for the convergence of the next subgraph iteration. Unlike most of existing research that process each subgraph only once and then overwrite their updates [13], [14], we propose to
iterate each subgraph multiple times so as to fully exhaust the value of each subgraph for avoiding the redundant data transmission and unnecessary iteration. Guided by this principle, we can thus make an innovation of leveraging the powerful yet ‘limited’ GPU processing capability to accelerate the multi-time subgraph iteration, enabling the scale-up efficiency for heterogeneous graph systems.

In an effort to break the update limitation within the subgraph, we present to pipeline the iteration of subgraphs. Compared with CLIP [15] that iterates on a fixed subgraph loaded from disk, our pipelined subgraph iteration is novel with a maximum scope of the subgraph information propagation. Further, based on existing high-optimized GPU-based computing model (e.g., pull execution model), we observe that a large number of vertex computations do not necessarily contribute to a valid vertex updating, causing to waste a large amount of GPU processing capacity. We propose a predictive vertex updating, aiming at efficiently identifying these vertices and further eliminating the unnecessary computations on them for better supporting pipelined subgraph iteration.

We compare Seraph with two state-of-the-art heterogeneous graph systems. Our results on a wide variety of real-world graphs demonstrate that Seraph outperforms Graphie [13] and Garaph [14] by 5.42x and 3.05x, respectively. In particular, Seraph can be significantly scaled up over existing heterogeneous graph systems. In addition, we compare Seraph with other large-scale graph processing solutions, revealing that Seraph can also achieve impressive performance in comparison to state-of-the-art CPU-based (i.e., Ligra [6]) and distributed graph processing systems (i.e., Gemini [4]).

The rest of this paper is as follows. Section 2 gives the background and motivation. We present pipelined iteration in Section 3. Section 4 elaborates predictive vertex updating. Section 5 shows the results. We survey the related work in Section 6. Section 7 concludes the work.

II. BACKGROUND AND MOTIVATION

In this section, we first give a brief introduction to GPU-accelerated heterogeneous architecture, followed by a motivating study regarding the inefficient data transmission of existing heterogeneous graph systems for large-scale graph processing, finally motivating our approach.

A. Heterogeneous Architecture

There emerge various heterogeneous architectural designs. Some are dedicated to performance improvement [16], some for energy reduction [17]. Some take both into consideration [18]. This paper has focused on GPU-accelerated heterogeneous architecture.

Figure 1 illustrates a typical GPU-accelerated heterogeneous architecture, which integrates the hardware advantages of both host side (with larger host memory) and GPU accelerator (with stronger computing ability). A GPU accelerator generally consists of multiple streaming multiprocessors (SMXs), each of which includes hundreds of cores.

In comparison to the high-speed internal bandwidth (e.g., ∼700GB/s for NVIDIA Tesla P100) of GPU cores accessing global memory, GPUs are generally connected to the host side with the relatively slow interface. For instance, the transmission bandwidth between CPU and GPU via PCIe Express 3.0 lane connection can be limited as slowly as ∼12GB/s in practice [19]. This significant gap may severely suppress the performance potential of heterogeneous architecture if the data is frequently transferred [11], [13].

Though there are a number of transmission interfaces that can provide higher interconnect bandwidth (e.g., Intel QuickPath Interconnect with 25.6GB/s, NVLink high-speed interconnect with 160GB/s), this paper focuses on PCIe interconnect since it is more common in the current commodity market.

B. Inefficiency of Existing Heterogeneous Graph Systems: A Motivating Study

In an effort to leverage the hardware advantages of heterogeneous architecture for large-scale graph processing, existing heterogeneous graph systems generally divide the entire graph data into subgraphs [13]–[15]. The CPU offers the graph data to the GPU in the form of subgraph. Once each subgraph is consumed, the GPU requests to process the next subgraph which is transferred from the host memory to the GPU global memory.

As a representative of state-of-the-art heterogeneous graph processing system, on the basis of basic subgraph iteration above, Graphie [13] additionally has two highlighted optimizations. First, it is ensured that only those subgraphs that have at least one active edge or vertex can be transferred.
Second, an asynchronous runtime to reuse the transferred subgraph at the next iteration. As a consequence, Graphie can partly reduce an amount of data transmission. Nevertheless, the potential of GPU-accelerated heterogeneous architecture can be still limited since their approach may be still inefficient to handle many graph algorithms (e.g., CC) where almost all subgraph can be active in the first few iterations. As a consequence, the majority of subgraphs still need to be transferred to the GPU within the iteration.

We investigate the performance characterization of Graphie’s cached subgraph iteration on three well-known graph algorithms as the number of available GPU SMXs is increasing in Figure 2. More details regarding the experimental settings can be found at Section V-A. Offering more SMXs is helpless for improving the efficiency of large-scale graph processing. As is known, current mainstream GPU accelerators usually have far more than 4 SMXs. There remains a significant gap between the low data transmission efficiency and high GPU computing capability for large-scale graph processing. Unfortunately, existing heterogeneous graph systems rarely respond to this challenge easily for providing the scale-up efficiency.

Under the premise of a fixed transmission bandwidth, one viable approach to improving the CPU-GPU transmission efficiency is to increase the bandwidth utilization with the well-organized subgraph data. Nevertheless, there still remain two significant defects at least for this approach. First, on account of the random memory accesses, graph processing usually behaves poor data locality [12], [20]. It is extremely difficult, if not impossible, to prepare a high-quality subgraph that can be fully used for each iteration. Second, even if we can identify such high-quality subgraph, it is also difficult to gather these data in a cost-efficient manner at runtime since graph re-organization is a well-known time-consuming process that may take more time than graph processing for many graph datasets [21], [22].

C. Our Observations

This work aims at reducing the impact of slow subgraph transmission on the high-performance graph computing. Instead of expensively preparing the high-quality subgraphs, we have the key insight that the transmission efficiency of heterogeneous graph systems can be greatly improved by making full use of the value of each subgraph, backed up by our observations as follows.

Observation 1: Each subgraph in one iteration has much useful information that serves to the subgraph processing in the next iteration.

Each subgraph is structured with many vertices that are associated via the edges. It has been observed that many graph algorithms (e.g., BFS and SSSP) are incremental iteration method that can greatly benefit from iterating each subgraph multiple times [15]. More specifically, in one iteration, the information of a given vertex can be only propagated to its neighboring vertices, which still full short in informing all the vertices that may be involved for graph processing. In order to handle all these vertices, there may need more iteration times to do so, causing to repeatedly load a large amount of redundant graph data. This inspires us of multi-time subgraph iteration to fully exploit its potential value for improving the transmission efficiency.

Observation 2: Multi-time subgraph iteration enables to improve the efficiency of heterogeneous graph processing by exploiting GPU processing capability.

Multi-time subgraph iteration enables the propagation and sharing of information among the subgraphs with multiple hops. This also provides a possibility to reduce the original task with multiple iterations into one iteration. Thus, a new question is how to efficiently iterate each subgraph multiple times, which largely depends on the GPU processing capability. That is, we are allowed to further improve the performance of heterogeneous graph systems through exploiting GPU processing capability.

With the above-discussed observation, the transmission inefficiency problem of heterogeneous graph systems is transferred into a problem of enhancing CPU-based graph processing, for which, a wide of highly optimized techniques can be directly leveraged [8].

III. PIPELINED SUBGRAPH ITERATION

Guided by Observation 1, we present a pipelined multi-time subgraph iteration, which is designed to make sure the information of each subgraph iteration can be propagated to a larger scope so that the value of each subgraph can be totally exhausted.

A. Preparation

We start by introducing the requisite preliminaries for the pipelined subgraph iteration.

Subgraph Organization Unlike the edge list organization used in prior work [13], Seraph uses more compact graph structure that manages to minimize the transmission data demand as much as possible. Specifically, Seraph only needs to transfer the Compressed Sparse Column (CSC) to the GPU. In order to process large-scale graph, CSC structure will be cut into much smaller CSC pages which include the a set of continuous vertices with the corresponding incoming edges. Compared to the page structure in [14], our data structure is more compact for omitting the destination vertex index array.

Heterogeneous Execution Model Seraph processes large-scale graphs with a heterogeneous execution model. Similar with prior work [4], [6], [14], [23], Seraph also adopts a density-aware model, which is efficient with less time-consumption. Our heterogeneous execution can be described as follow. 1) In the sparse stage, we compress the active frontier and perform the sparse updating on the CPU with push-based execution model, so as to make sure GPU
B. Pipelined Subgraph Scheduling

The key idea behind our multi-time subgraph iteration is to pipeline the subgraph scheduling for the purpose of maximizing the information propagation of each subgraph to other subgraphs. Streaming topology is widely used in network optimization [24], and also in storage hierarchy optimization. CLIP [15] presents a disk-based "loaded data reentry (ldr)" streaming topology to squeeze out the value of loaded data. It has the following features: 1) process the subgraph more than once to reduce the total iteration times; and 2) subgraph data is loaded by sequential order to maximize disk IO bandwidth. Nevertheless, the loaded data reentry this way is prone to causing much redundant computation. The underlying reason accounting for this is because the information can not propagate among multiple subgraphs for timely interaction, and block the convergence of the entire graph.

Figure 3 shows an example of CLIP’s subgraph iteration with BFS on the given graph based on pull model. Considering the similarity, we only give a part of the entire graph. With the loaded data reentry optimization, we can find that BFS has no sign of convergence, but simply process subgraph one by one has converged quickly with the same schedule times. It is because the subgraph $A$ and subgraph $B$ can not share the update within the multi-reentries. More specifically, the update of $A$ can notify $B$ after $A$ has finished ($1 \rightarrow 4$), but the update of $B$ can not feed back to subgraph $A$ within the multi-reentry ($4 \rightarrow 2$), since subgraph $A$ has been discarded from this iteration. That is, the graph partition destroys the sub-structure among subgraphs, further blocking information propagation between multiple subgraphs.

Pipelined Subgraph Iteration We therefore present a novel subgraph iteration in a pipelined fashion for maximizing the propagation scope and rebuild the substructure among subgraphs. Our pipelined subgraph iteration method is shown in Figure 4. For facilitating the descriptions, considering GPU memory space can maximally hold four subgraphs, and we assume that the GPU can process three subgraphs and transfer one subgraph at the same time. Then kernel execution on subgraph $A, B, C$ can be overlapped with the transmission of subgraph $D$. We could treat subgraph $A, B$ and $C$ as a “super subgraph” since we process the vertices and edges in $A, B, C$ concurrently. When the kernel finished, subgraph $D$ has just been copied from the CPU, we then launch kernel to process subgraph $B, C, D$ and transfer subgraph $E$ repeatedly. Afterwards, the space occupied by subgraph $A$ is overwritten.

However, to maximize to substructure and fully overlap the data transfer with kernel execution, double-buffer based schedule workflow can be viewed in figure. Similarly, this method also load the graph once but process more than once, furthermore process the subgraph $AB$ three times and transfer subgraph $CD$ at the same time.

But our approach is more efficient for two reasons. First, we can offer a larger scope for information propagating. As figure 4 shows, the $D$ can even reconstruct substructure with $B$ and $E$, enlarging the information propagate scope to at most five subgraphs in the example compared with two in $ldr$ used in CLIP. Second, pipelined subgraph iteration further allow us to make a fine-grained scheduling, and we introduce as follows.

Example mentioned above assumed that GPU can process three subgraphs and transfer one subgraph at the same time. However, it is hard to reflect most of real-world situations since the execution time is constantly change, and there are many idle slots (transmission wait for computation or computation wait for transmission). In order to fill the idle slots between subgraph transmission and computation, Seraph minimizes the effect by a more fine-grained scheduling. Specifically, if the transmission thread finds that the transmission task has finished, but the computation has not finished yet, it will check if the kernel executed on the first few subgraph (the subgraphs will be discarded from the execution set immediately) has finished. If so, the thread will transfer the next subgraph to GPU, and overwrite the space occupied by the finished kernel. On the contrary, if the computation has finished but the transmission not, Seraph will select one subgraph in the finished kernel set to reentry once. This method mainly benefit from much smaller
subgraphs used in Seraph, and offer an incremental model to fill the idle slot as much as possible. 

Remark With subgraph iteration pipelined, we enable to use more information within each iteration for the fewer iteration numbers and faster convergence speed. The vertex update in one subgraph could spread to a large scope rebuilt by multiple smaller subgraph, making the information spreading more sufficient.

IV. Predictive Vertex Updating

Guided by Observation 2, this section mainly presents predictive vertex updating. We first review the characteristics of vertex updating with pipelined subgraph iteration. In accordance with characteristics analysis, we further present two efficient solutions of vertex updating to enhance the GPU processing capability.

A. Characteristics Analysis of Vertex Centric Updating

In this work, we follow the pull-based execution model on GPU which involves few data races, so that the data-parallel potential of GPU can be fully exploited. However, almost all existing systems with density-aware optimization perform the pull update attempt on all vertices. Nevertheless, though a part of vertices can be updated, we note that many of other vertices updating attempts have failed, leading to a large number of unnecessary computations.

Figure 5 shows an example of connected component with a label-propagation-based method on a given graph. The algorithm follows a synchronous execution model. We find that there may involve two unnecessary vertex updating situations. 1) The vertices have converged before the iteration (e.g. vertex 5 has converged after the first iteration). 2) The vertices are far from the global minimal label but converged with local minimal label (e.g. vertex 6 will get label 1 at the 5th iterations, but with label 2 during the 1st~4th iteration). Though these updates make no benefit for graph convergence during the iteration, but it is difficult to predict whether update attempt will fail or succeeded in this iteration.

Accurately identifying these unnecessary vertex computations are the key to enhance the GPU processing capability. In the following, we next introduce two efficient yet accurate solutions to predict whether a vertex value would be useful in making a valid vertex update at this iteration towards two situations above.

Table I: Strong convergence condition on BFS, WCC and SSSP. \( \text{Value}[i] \) means the vertex value for various graph algorithms, e.g., the depth for BFS and distance for SSSP. \( ccs\text{size}[v] \) represents the number of vertices labeled with \( v \).

| Condition       | Definition                                                                 |
|-----------------|----------------------------------------------------------------------------|
| BFS \( \text{value}[i] \leq k \) | \( k \) is the iteration times                                              |
| CC \( \text{value}[i] \leq s \) | \( s = \min \{\text{ccsize}[v](k-1) \neq \text{ccsize}[v](k-2) \} \) |
| SSSP \( \text{value}[i] \leq l \) | \( l = \min \{\text{value}[v](k-1) \neq \text{value}[v](k-2) \} \) |

B. Strong Prediction Condition

We find that many vertices could be judged as converged by simply according to the vertex value. For example, vertex 5 in figure 5 is converged just after the first iteration, because it has already gotten the smallest label of 1 in the given graph. As the iteration number is increasing, the vertex value changes in a monotonous pattern. We thus list a strongly predictive solution in Table I. We prove the condition of convergence separately. And note that the condition is suitable for asynchronous update, naturally suitable for synchronous model.

For BFS, we can conclude that the vertex has converged if the value of vertex \( i \) is smaller than the iteration times \( k \). It is easy to prove that if there is a shorter unweighted path \( r (r < k \text{ and } r < \text{value}[v]) \) from the source vertex, the path must have been found in \( r \) times iteration. Though the top-down/down-top based BFS also benefits from omitting the update for traversed vertices, but it is not suitable for asynchronous execution model, since the vertex could be traversed with a longer path for the much faster asynchronous update, but not converged actually.

For CC, we can conclude that the vertex has converged if the vertex has been chosen into a converged component. More specifically, the component size will not increase if the component root (minimal vertex in the component) has packed all the vertex into the component, and we call the component has converged. For a vertex, the optimal situation is finding an component not converged yet. Thus the minimal label could be \( s \), the root for the component not converged yet. So the vertex must have converged if the vertex label is smaller than \( s \). A special but really useful case is that the vertex has got the minimal label of the whole graph, then can be concluded converged immediately.

For SSSP, we focus on weighted graph without negative edge. We can conclude the vertex has converged if the vertex has found a path shorter than the latest shortest path in the last iteration. It is easy to prove, since the vertex’s distance to the source vertex is based on the more closer vertex on the shortest path. It is impossible to get a smaller distance than the closest one at the last iteration.

The strong predictive vertex updating method can judge the vertex converged or not directly from the vertex value, regardless of the graph topology. And we only need to compared the vertex value with the threshold, then if the
Figure 6: Weakly predictive vertex updating state-transition diagram. We let the vertices with status 2, 3, 4 to be not updated. If there is a valid update, we return to initial state 0 by following the red arrow, otherwise following the black arrow.

vertex value satisfy the condition, the cost for computation and the memory access on this vertex can be all reduced.

C. Weak Prediction Condition

Though strong prediction can reduce the unnecessary vertex update directly, however, in many cases it is difficult to design a strong converged condition such as data-driven PageRank (PageRank-delta). We thus present an alternative weak predictive technology based on the updating history.

Inspired by using the branch history to predict the branch output [23], we also try to predict the vertex update result with the update history. Firstly we try to find the update patterns for graph processing, so we keep a record of the update history for each vertex with various graph datasets and algorithm (note that we only record the iteration result processed on GPU side with asynchronous execution model). For example, a vertex update history is [1, 1, 0, 1, 0] indicating that the vertex value changed at the 1, 2, 4 iteration, but the update attempt in iteration 3 and 5 failed, indicating the vertex value do not change at these two iterations. And we find an interesting pattern: if the vertex value has changed before, but the value no longer change with the two subsequent iteration, the vertex can be concluded as converged with a high likelihood (91.3% for SSSP on our dataset, and about 98% for CC and BFS). We conclude this pattern with a lot of off-line analysis.

Based on the afore-mentioned statistical regularities, we further present a state-transition diagram based solution to guide the vertex updating with low overhead. And the diagram should match the statistical regularities to predict the vertex update result effectively.

Figure 6 shows the state-transition diagram designed for our system. There are six states used in the diagram. The vertices with state 0, 1, 5 indicate that this vertex should make an update attempt during this iteration, and the update attempts for the vertex with state 2, 3, 4 will be canceled during the iteration. At the beginning, all vertices are initialized as the state 0. After the first iteration, the vertex state will get into 5 (dormancy candidate state) if the update attempt failed, or it returns back to 0 if success. The vertices with state 5 will get into 3 (dormancy state) if the update failed again. This design ensures that all vertices with state 3 has failed at least two times and we can assume the vertex has converged with a high likelihood, which matches with the statistical regularities mentioned above. Vertex with state 3 will be dormant and do not make an update attempt at the next two iterations, then memory access and computation can be canceled on the vertices with these status. To avoid the vertex getting into dead state and no longer attempt to update, we add state 1 to make an attempt after the two round of dormancy iterations, but if the update failed again, the vertex will get into state 4 as a punishment.

This method can be integrated into the graph engine naturally. Since Seraph follow a vertex-centric programming model, we just need to allocate a status word for each vertex and initialize it to state 0, then the status can be driven by checking the vertex value changed or not during this iteration, or just turn into the next state when it is in dormancy state. And the state-transition diagram is defined by users as a input for graph engine, making the graph algorithm unmatched with this diagram can also benefit from this method by using other diagram. And though the weak predictive vertex update may be false, we will make an all-pull update to retrieve the active vertices within the dense to sparse switch.

Discussion We find that the BFS, CC with strong predictive update method is really efficient, but SSSP can get a better efficiency with weak predictive method. We analyze the update history and find that SSSP with strong predictive vertex update can only reduce a small amount of useless update accurately, but weak predictive update method can remove considerable amount of useless update though may be false with low possibility.

With the predictive vertex updating, many useless update can be identified and further reduce the amount of computation. And these saved computation can be used to squeeze out the real value of loaded data, improve the system efficiency directly.

V. Evaluation

In this section, we evaluate the efficiency of Seraph by answering the following four research questions:

- **RQ1**: How efficient is Seraph compared to existing heterogeneous graph systems? [Section V-B]
- **RQ2**: How does Seraph scale with varying graph sizes and processing capability? [Section V-C]
- **RQ3**: How effective are pipeline subgraph iteration and predictive vertex updating? [Section V-D]
- **RQ4**: How well is Seraph advantageous over other state-of-the-art large-scale graph parallel processing systems? [Section V-E]

A. Experimental Setup

**Testbed** All tests are performed on a machine where the host side is equipped with an Intel 10-core Xeon E5-2650v3@2.40GHz, 64GB main memory. The GPU acceler-
ator is NVIDIA GTX980 (with 16SMXs, 2048 cores and 4GB GDDR5 memory), which is connected to the host side via PCI Express 3.0 lanes operates at 16x speed. The transmission bandwidth for asynchronous memory copy between CPU and GPU is around 11GB/s.

**Methodology** Graphie [13] and Garaph [14] are the most related GPU-accelerated heterogeneous systems to our Seraph. Graphie tracks the active subgraph with a compact data structure, and only transfers the active subgraph from the CPU to GPU to reduce the useless data transfer. It gives the priority to process the subgraph buffered on GPU after last iteration. Garaph dispatches the subgraphs between the CPU and GPU during the dense stage, and processes the subgraphs on the CPU side during the sparse stage.

Unfortunately, neither Graphie nor Garaph is open sourced. For making a comparison with them, we can thus only reference the experimental results reported in their work (just as previous work has also done [13], [20]), and evaluate Seraph with the same graph datasets that have been used. Table II depicts the detailed machine specification of Seraph against those for Graphie and Garaph. We note that Seraph basically has the worst configuration but gains the best performance as will be discussed in the subsequent experiments.

**Graph Algorithms** We evaluate Seraph using three well-known graph traversal algorithms: 1) Breath-First Search (BFS) for traversing the graph hop by hop so as to compute the distances of all vertices from a specific vertex; 2) Connected Components (CC) for finding a maximal number of subgraphs where any two vertices can be connected via a chain of paths; and 3) Single-source Shortest Path (SSSP) for finding a path of a given vertex to every vertex such that the sum of the weights of their constituent edges is minimized. As discussed in Section 4.3, we evaluate BFS and CC with strong predictive vertex updating optimization. SSSP is evaluated with weak predictive vertex updating optimization.

**Graph Datasets** We benchmark the graph algorithms with a variety of graph collections, including: 1) 12 real-world graphs (coming from Stanford Large Network Dataset Collection [1] and Laboratory for Web Algorithmics [2]); and 2) 6 large synthesized graphs (generated by the RMAT tool). Table III depicts the graph collections that we have used for the aforementioned comparison with Graphie and Garaph, and further evaluation.

### RQ1: Efficiency

We first evaluate the efficiency of Seraph in comparison to two state-of-the-art GPU-accelerated heterogeneous systems: Graphie and Garaph. In an effort to make a fair comparison, like Graphie and Garaph, our results take both data transfer and kernel execution into account.

---

**Table II: Detailed machine specifications that have been used in Garaph, Graphie and Seraph**

| Spec.          | Graphie | Garaph | Seraph |
|----------------|---------|--------|--------|
| CPU type       | NVIDIA Titan Z | NVIDIA GTX1070 | NVIDIA GTX980 |
| On-board memory| 6GB GDDR5 | 8GB GDDR5 | 4GB GDDR5 |
| Internal bandwidth | 288GB/s | 256GB/s | 224GB/s |
| CUDA cores     | 2688    | 1920   | 2048   |
| Host memory    | E7-4830 v3 | E5-2650 v3 | E5-2680 v3 |
| Graphie        | 256GB DDR3 | 64GB DDR4 | 64GB DDR3 |
| Garaph         | 6GB GDDR5 | 4GB GDDR4 | 4GB GDDR4 |
| Seraph         | 6GB GDDR5 | 4GB GDDR4 | 4GB GDDR4 |

**Table III: Graph datasets used in our evaluation**

| Dataset       | | |
|---------------|---|---|
| dataset       | $|V|$ | $|E|$ |
| cage15        | 5.1M | 99.1M |
| kron-500      | 2.1M | 182.1M |
| nlpkkt160     | 8.3M | 221.1M |
| orkut         | 3.1M | 117.2M |
| uk-2002       | 18.5M | 298.1M |
| twitter-2010  | 61.6M | 1,468.4M |
| friendster     | 124.8M | 1,806.1M |

We make $22 < k < 29$ to create different graph sizes.

Compared with Graphie Table IV shows the detailed comparative results. It is worth noting that, for Graphie, Garaph and Seraph, twitter and friendster can not fit into their GPU global memory. Seraph provides a considerable performance benefit with 4.2x on average due to our pipelined subgraph iteration that has less requirement on data transmission.

For the small graphs, all graph data can be fit into the GPU global memory. In this case, though the pipelined subgraph iteration might be considered helpless, our predictive vertex updating can still be efficacious in enhancing in-memory computing. As we can see, Seraph significantly outperforms Graphie for almost all cases. For instance, BFS with nlpkkt160 can obtain up to 24.64x speedup. We should note that, for SSSP with cage15 and nlpkkt160, Seraph has little performance improvement that is because nlpkkt160 show up meshwork property which is more suitable for a push-based SSSP algorithm [6].

Compared with Garaph Table V depicts the comparative results. Since Graphie does not provide the results regarding BFS, we thus test Seraph using CC and SSSP only. As a consequence, we can see that Seraph has a better performance than Garaph for all graph datasets, with up to 5.6x speedup for in-GPU-memory graphs and 2.6x speedup for out-of-GPU-memory graphs. Note that GTX 1070 used in Garaph has a better configuration than Seraph’s, especially with double memory capacity with ours. There are two reasons for Seraph outperforms Garaph. Though Garaph reduces a part of I/O during the dense stage, but GPU is more powerful than CPU, most subgraphs also have been transferred to the GPU, but Garaph does nothing to cover the gap between computation and data transfer. Second Garaph adopts all-pull method to process the edge during the dense stage, which includes much useless computation, and Seraph reduces these computation with the predictive vertex update.
Table IV: Seraph vs. Graphie

|       | Graphie Seraph speedup | Graphie CC speedup | Graphie SSSP speedup |
|-------|------------------------|--------------------|----------------------|
| cage15 | 0.63 0.095 6.63x       | 0.23 0.066 3.48x   | 0.24 0.54 0.44x     |
| iron_g500 | 0.59 0.047 12.55x   | 0.48 0.057 8.42x   | 1.67 0.135 12.37x  |
| nplike160 | 6.11 0.248 24.64x  | 8.02 0.73 14x     | 7.03 8.42 0.83x    |
| orkut | 0.21 0.067 3.13x      | 0.26 0.088 2.95x   | 0.6 0.24 2.5x      |
| uk-2002 | 4.3 0.23 18.7x       | 5.04 0.35 14.4x    | 11.73 0.85 13.8x   |
| twitter | 5.42 1.06 5.11x    | 4.21 1.09 3.86x    | 14.67 4.8 3.06x    |
| friendster | 16.44 3.8 4.32x  | 12.46 2.3 5.42x   | 29.24 7.07 4.14x   |

Table V: Seraph vs. Garaph

|       | Garaph Seraph speedup | Garaph CC speedup | Garaph SSSP speedup |
|-------|------------------------|--------------------|----------------------|
| uk2007@1M | 0.14 0.038 2.92x    | 0.48 0.11 4.36x   |                      |
| uk2014-host | 0.17 0.09 1.89x  | 0.57 0.236 2.42x  |                      |
| enwiki-2013 | 0.28 0.05 5.6x     | 0.7 0.174 4.02x   |                      |
| gsh-2015-tpl | 1.21 0.42 2.88x  | 4.32 1.8 2.4x     |                      |
| twitter | 3.32 1.09 3.05x    | 12.75 4.8 2.66x   |                      |
| sk-2005 | 4.47 4.01 1.12x   | 18.13 11.13 1.65x |                      |

Though Garaph reduces the data transfer during the sparse iteration, Seraph also reduces these IO overhead with a more efficient push-based method.

C. RQ2: Scalability

We investigate the scalability of Seraph by: 1) controlling the number of SMXs available for graph processing, and 2) adjusting the graph size that varies from in-memory scale to out-of-memory scale.

Scaleability with varying SMXs Figure[2]a depicts the performance characterization between Graphie's subgraph iteration and ours by benchmarking entire uk2007 dataset (with 105 million vertices and 3.7 billion edges). With our technical highlights of pipelined subgraph iteration and predictive vertex updating, it is revealed that, for all three graph algorithms, Seraph can significantly improve the performance of graph processing over Graphie as fed with more GPU SMXs. For instance, for BFS, when the number of SMXs reaches at 4, the performance of Graphie will be saturated. In contrast, Seraph can continue offering a near-linear performance improvement. CC and SSSP have the similar observation.

We should note that Seraph may also enter into saturation. For BFS, the benefits will stop when the number of SMXs reached at 11. We guess one of underlying reasons may lie in the limited GPU internal bandwidth between SMX and global memory. Figure[2]b further shows the performance characterization of pure in-memory computing using the small graph enwiki-2013 that can fit into the global memory. We find that a similar observation has occurred. This yields a conclusion that the latter saturation has nothing to do with the CPU-GPU transmission efficiency. Coping with the internal inefficiency of GPU for graph processing can be interesting future work, which is beyond the scope of this paper.

Scaleability with varying graph sizes Figure[3] depicts the performance characterization using different strategies with different scale of RMT dataset where the edge scale is 16x larger than vertex's. Note that GTX980 has 4G Bytes global memory. BFS and CC run on unweighted graphs, and hence, they can load the unweighted RMT with 1 Billion edges at most into GPU global memory. SSSP works on weighted graph that can have 512 million edges at most into GPU global memory.

Overall, the throughput (i.e., GTEPS) is high when the graphs can be load into global memory. Further, it will go down dramatically when the graph comes to the critical point that is out of global memory because of the CPU-GPU data transmission. To be specific, predictive vertex updating can introduce more benefits, especially for the in-memory situations. As for large-size graphs, we can still find that our pipelined subgraph iteration and predictive vertex updating contribute to considerable performance benefits. Note that the host memory will run out for SSSP on RMT28 with 4-billion edges, which needs at least 64GB to store the weighted graph data.

D. RQ3: Effectiveness

We also evaluate the effectiveness of pipelined subgraph iteration (psi) and predictive vertex updating (pvu).

Breakdown Figure 9 shows the breakdown results. We conduct the test on four large graph datasets where TW is twitter2010. RMT represents RMAT27 and UK indicates the entire dataset of uk2007. The baseline is the traditional subgraph processing without multiple iterations [13]. In comparison to baseline, we can see that psi can offer up to 1.86x speedup for CC on uk2007, and especially for large-scale graph, because the subgraph cache optimization is useless when the graph size is much larger than memory size. The predictive vertex update optimization can offer up to 1.79x speedup on twitter, and offer 1.46x speed up on average. The total benefit can get from these two technology is up to 2.86x speed up for CC on friendster, and offer 2.11x speed up on average.

Effectiveness of pipelined subgraph iteration To demonstrate the effectiveness of psi, we also test naive reentry subgraph iteration (rsi) that have been used in
CLIP [15]. For CLIP, we have set the most suitable value of maximum reentry times (MRT) for BFS, CC and SSSP to be 3, 2 and 2, respectively. More details regarding how to set the reasonable value of MRT can be found in [15].

Figure 9 lists the detailed results. We can see that rsi introduces a slight performance improvement in comparison to the baseline, with at most 34% improvement for CC on uk2007 as an example, and only 10.4% on average. This is because rsi adopts updating the subgraph over and over again, further causing much redundant computation. In comparison to rsi, psi cause less redundant computation due to the pipeline schedule method. As a result, we can get up to 86% performance improvement for CC on uk2007. The averaged performance improvement is by 44.4%.

**Effectiveness of predictive vertex updating** Relative to strong pvu that can definitely reduce the useless work, the weak pvu has more complex relations. Thus, we choose SSSP to evaluate the effectiveness of pvu for better demonstration.

Figure 10a) counts the number of vertices for each status during each iteration. It shows that the number of vertices that involve a data update (i.e., with status 0, 1 and 5) is significantly reduced as iteration goes deeper. That is, there is a large amount of redundant computation (i.e., with status 2, 3 and 4) that does not lead to a valid vertex updating is increasing during each iteration. Figure 10further sums up the number of vertices (with status 0/1/5) as our label, and those (with 2/3/5) as as redundant label. Real indicates the number of vertices that involve the valid updating during each iteration. Compared to the total amount of computation in prior work [6], [14], our optimization can basically reflect the real situation, and also, dramatically reduce their redundant computation for performance speedup.

**E. RQ4: vs. Other Advanced Large Scale Graph Processing Solutions**

We finally compare Seraph with other state-of-the-art solutions for large-scale graph processing: 1) Ligra [6], a host-memory based solution; and 2) Gemini [4], a distributed solution. Since Ligra is memory consuming [14], we thus extend to use a node with 256GB memory for testing it. Gemini is set on a 4-node cluster. Each node has the same
configuration as the one for Ligra.

Figure 11 depicts the results. In comparison to Ligra, it can be seen that Seraph is much efficient for CC and SSSP with up to 3.9x speedup. Heterogeneous solution with an integration of specialized accelerator can provide a great potential over traditional shared memory solution. Even compared to Gemini (with more computing and storage resources), Seraph can also obtain the comparable results. For instance, our results for CC on all datasets are superior to the ones using Gemini, with up to 1.64x speedup. Though Gemini is a good scale-out solution, heterogeneous solution can also offer comparable efficiency, mainly thanks to less network communication and fewer redundant computations.

Note that both Ligra and Gemini show a better efficiency than Seraph for BFS. The underlying reason lies in some specialized optimization they have used. To be specific, vertex update will break early during the dense iteration if the vertex has found a neighbor vertex has been traversed. So the computation for BFS is much sparse than CC and SSSP. We should also note that such tailor-made optimization can be also implemented into Seraph framework for accelerating BFS.

VI. RELATED WORK

Heterogeneous Graph Systems Heterogeneous architecture has integrated the advantageous resources of different devices for satisfying different demands of graph processing. A wide spectrum of efforts have been put into developing specialized graph accelerators via single- [26]–[28] or multi-FPGAs [29] for energy-efficiency purposes. There also emerge a number of GPU-accelerated heterogeneous graph systems [8], [10], [23], [30] for supporting high-performance large-scale graph processing. Graphie [13] transfers the active subgraph to GPU and reuses the cached subgraph in the next iteration to reduce the I/O amount. Garaph [14] streams edge data asynchronously to the GPU for graph processing.

As discussed before, these prior graph systems may still fall into limited performance when handling large-scale graphs due to their inefficient data transmission. This paper first (to our best knowledge) proposes to leverage multi-time subgraph iteration to break through this limitation, enabling the scale-up efficiency of heterogeneous graph systems.

Distributed Graph Systems A large amount of research seeks help from distributed deployment that can aggregate more resources than single machine for processing large-scale graphs. The primary task of distributed graph systems is to obtain well-cut graph partitions [2]–[4], [31], [32] so as to minimize the communication across machines. A few recent studies use the emerging high-speed network (e.g., RDMA) to reduce the communication overhead [33], [34]. Gemini presents a series of adaptive runtime optimizations with sparse-dense switching, locality-aware and NUMA-aware features, enabling an attractive scale-out efficiency.

In comparison to these distributed designs, it is verified in our evaluation (Section 7.5) that heterogeneous solutions, in spite of involving fewer resources, can be also or even more promising in practice for large-scale graph processing because of less communication cost and fewer redundant computations.

Disk-based Graph Systems There also exist many disk-based systems to support large-scale graph processing. GraphChi [5] proposes parallel sliding windows to lead only non-sequential accesses to the disk. GridGraph [21] uses 2-level hierarchical partitioning to reduce the I/O amount. TurboGraph [35] presents a pin-and-slide model to fully exploit the multicore and I/O parallelism. Since the significantly low disk-to-memory bandwidth, disk-based graph systems are orders-of-magnitude slower than heterogeneous solutions.

In-Memory Graph Systems For in-memory graph processing, the graph data only needs to be copied at the beginning. Once data is ready, the processors can process the graph without any data transmission until finished [6]. [7], [22]. Usually, in-memory graph systems, such as Gunrock [8], can provide orders of magnitude performance improvement over heterogeneous implementations. However, limited to global memory, existing dedicated accelerators (e.g., GPU) can not store real-world graphs (with more than billions edges) [6], [36]. Considering the inefficient data transmission, the potential of accelerators is also significantly underutilized. On the contrary, combined with pipelined subgraph iteration, we further present predictive vertex updating to better exploit the GPU processing capability for enhancing the performance of heterogeneous graph systems.

VII. CONCLUSION

There remains tremendously challenging for scaling up the performance of heterogeneous graph systems due to the well-known interconnect transmission inefficiency. To cope with problem, by iterating each subgraph multiple times, it is observed that the heterogeneous graph systems can further obtain an improvable performance by enhancing GPU processing capacity.

With this guideline, we develop Seraph integrated with two technical innovations. First, we present a pipelined subgraph iteration to maximize the information propagation of each subgraph to other ones for fully exhausting its value. Second, we propose two efficient vertex updating solutions to predict unnecessary vertex computations for better supporting pipelined subgraph iteration. Our results demonstrate that Seraph outperform the state-of-the-art heterogeneous graph systems Graphie and Garaph by 5.42x and 3.05x, respectively. Seraph can be also scaled up over existing heterogeneous graph system. Our comparative results also reveal that Seraph can achieve impressive performance in comparison to state-of-the-art CPU-based (i.e., Ligra) and distributed graph processing systems (i.e., Gemini).
[26] J. Zhang, S. Khoram, and J. Li, “Boosting the performance of fpga-based graph processor using hybrid memory cube: A case for breadth first search,” in Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA), 2017, pp. 207–216.

[27] S. Zhou, C. Chelmis, and V. K. Prasanna, “High-throughput and energy-efficient graph processing on fpga,” in 2016 IEEE 24th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), 2016, pp. 103–110.

[28] H. Jin, P. Yao, X. Liao, L. Zheng, and X. Li, “Towards dataflow-based graph accelerator,” in 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), 2017, pp. 1981–1992.

[29] G. Dai, T. Huang, Y. Chi, N. Xu, Y. Wang, and H. Yang, “Foregraph: Exploring large-scale graph processing on multi-fpga architecture,” in Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA), 2017, pp. 217–226.

[30] C. Hong, A. Sukumaran-Rajam, J. Kim, and P. Sadayappan, “Multigraph: Efficient graph processing on gpus,” in Parallel Architectures and Compilation Techniques (PACT), 2017 26th International Conference on. IEEE, 2017, pp. 27–40.

[31] R. Chen, J. Shi, Y. Chen, and H. Chen, “Powerlyra: Differentiated graph computation and partitioning on skewed graphs,” in Proceedings of the Tenth European Conference on Computer Systems. ACM, 2015, pp. 13–21.

[32] C. Avery, “Giraph: Large-scale graph processing infrastructure on hadoop,” in Proceedings of the Hadoop Summit. Santa Clara, vol. 11, no. 3, 2011, pp. 5–9.

[33] M. Wu, F. Yang, J. Xue, W. Xiao, Y. Miao, L. Wei, H. Lin, Y. Dai, and L. Zhou, “Gra m: scaling graph computation to the trillions,” in Proceedings of the Sixth ACM Symposium on Cloud Computing. ACM, 2015, pp. 408–421.

[34] J. Shi, Y. Yao, R. Chen, H. Chen, and F. Li, “Fast and concurrent rdf queries with rdma-based distributed graph exploration,” in Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation (OSDI), 2016, pp. 317–332.

[35] W.-S. Han, S. Lee, K. Park, J.-H. Lee, M.-S. Kim, J. Kim, and H. Yu, “Turbograph: a fast parallel graph engine handling billion-scale graphs in a single pc,” in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013, pp. 77–85.

[36] K. Zhang, R. Chen, and H. Chen, “Numa-aware graph-structured analytics,” vol. 50, no. 8. ACM, 2015, pp. 183–193.