Efficient Multi-GPU Computation of All-Pairs Shortest Paths

Hristo Djidjev and Sunil Thulasidasan
Los Alamos National Laboratory
Los Alamos, NM, USA
Email: {djidjev,sunil}@lanl.gov

Guillaume Chapuis, Rumen Andonov, and Dominique Lavenier
INRIA/IRISA, University of Rennes
Rennes, France
Email: {guillaume.chapuis,rumen.andonov,dominique.lavenier}@irisa.fr

ABSTRACT

We describe a new algorithm for solving the all-pairs shortest-path (APSP) problem for planar graphs and graphs with small separators that exploits the massive on-chip parallelism available in today's Graphics Processing Units (GPUs). Our algorithm, based on the Floyd-Warshall algorithm, has near optimal complexity in terms of the total number of operations, while its matrix-based structure is regular enough to allow for efficient parallel implementation on the GPUs. By applying a divide-and-conquer approach, we are able to make use of multi-node GPU clusters, resulting in more than an order of magnitude speedup over the fastest known Dijkstra-based GPU implementation and a two-fold speedup over a parallel Dijkstra-based CPU implementation.

I. INTRODUCTION

Shortest-path computation is a fundamental problem in computer science with applications in diverse areas such as transportation, robotics, network routing, and VLSI design. The problem is to find paths of minimum weight between pairs of nodes in edge-weighted graphs, where the weight $|p|$ of a path p is defined as the sum of the weights of all edges of p. The distance between two nodes v and w is defined as the minimum weight of a path between v and w.

There are two basic versions of the shortest-path problem: in the single-source shortest-path (SSSP) version, given a source node s, the goal is to find all distances between s and the other nodes of the graph; in the all-pairs shortest-path (APSP) version, the goal is to compute the distances between all pairs of nodes in the graph. While the SSSP problem can be solved very efficiently in nearly linear time by using Dijkstra's algorithm [1], the APSP problem is much harder computationally.

Two main families of algorithms exist to solve the APSP problem exactly: the first family is based on the Floyd-Warshall algorithm [2], while the second derives from Dijkstra's algorithm. The Floyd-Warshall approach consists in iterating through every vertex v_k of the graph to improve the best known distance between every pair of vertices (v_i, v_j) (see Algorithm 1). The complexity of this approach is $O(|V|^3)$, regardless of the density of the input graph. While the algorithm works for arbitrary graphs (including those with negative edge weights), its cubic complexity makes it infeasible for very large graphs.

Given that the Dijkstra algorithm solves the SSSP problem, it is possible to solve the APSP problem by simply running the Dijsktra algorithm over all source vertices in the graph (see Algorithm 2). When using min-priority queues, the complexity of this approach is $O(|E| + |V| \log |V|)$ for the SSSP problem, where V and E are the sets of the vertices and edges, respectively. For the APSP problem, the total complexity is thus $O(|V| * |E| + |V|^2 \log |V|)$, which becomes $O(|V|^3)$ when the graph is complete, but only $O(|V|^2 \log |V|)$ when $|E| = O(|V|)$, making this approach faster than Floyd-Warshall for sparse graphs.

While the All-Pairs Shortest Path problem regularly occurs in routing in transportation networks, it is also applicable to many other domains. It is the first step in obtaining several network measures that are of importance in domains such as social network analysis and bio-informatics. One such measure is the *betweenness centrality*, which is defined, for any vertex v, as the number of shortest paths between all pairs of vertices that pass through v. Betweenness is a measure of v's centrality (importance) in the network, and algorithms frequently use the centrality of the nodes in a network in order to compute the community structure of the network [3]. Furthermore, in several applications, the networks that need to be analyzed may have negative weights, and hence one needs an algorithm that solves the APSP problem for graphs with real (positive as well as negative) weights. In online social networks, for instance, negative weights may be used to indicate antagonism between two individuals [4] or even conflicts and alliances between two groups [5]. Causal networks in bioinformatics also use negative edges to represent inhibitory effects [6].

In this paper, we present an algorithm for solving the APSP problem for graphs with real weights that exploits the high degree of parallelism available in today's Graphics Processing Units (GPU). GPUs and other stream processors were originally developed for intensive media applications and thus advances in the performance and general purpose programmability of these processors have hitherto benefited applications that exhibit computational similarities to graphics applications, namely high data parallelism, high computational intensity, and data locality. However, many

Algorithm 1 Floyd-Warshall algorithm.

```
INPUT: A graph G(V,E), where V is a set of
    vertices
  and E a set of weighted edges between these
  vertices.
OUTPUT: The distance of the shortest path between
  any two pairs of vertices in G.

for each vertex v in V
    dist[v][v] = 0
end for
for each edge (u,v) in E
    dist[u][v] = w(u,v)  // the weight of the edge
    (u,v)
end for
for k from 1 to |V|
    for i from 1 to |V|
        for j from 1 to |V|
            dist[i][j] =
     min(dist[i][j], dist[i][k] + dist[k][j])
     end for
  end for
end for
return dist
```

Algorithm 2 Dijkstra's Single Source Shortest Path algorithm.

```
INPUT: A graph G(V,E), where V is a set of
    vertices
  and E a set of weighted edges between these
  vertices. A source vertex from V.
OUTPUT: The distance of the shortest paths between
  the source vertex and every vertex in V.

for each vertex v in V
  dist[v] = infinity
  previous[v] = undefined
end for
dist[source] = 0
Q = V
while Q is not empty
  u = vertex in Q with smallest distance in dist[]
  Q = Q\{u}
  if dist[u] = infinity
    break

  for each neighbor v of u in Q
    alt = dist[u] + dist_between(u, v)
    if alt < dist[v]
      dist[v] = alt
      previous[v] = u
      decrease-key v in Q
    end if
  end for
end while
return dist
```

theoretically optimal graph algorithms exhibit few of these properties. Such algorithms often use efficient data structures storing as little redundant information as possible, resulting in highly unstructured data and un-coalesced memory access making them less-than-ideal candidates for streaming processor manipulations. Nevertheless, given the wide applicability of graph-based approaches, the massive parallelism afforded by today's graphics processors is too compelling to ignore; current GPUs support hundreds of cores per chip and even future CPUs will be many core.

Our approach exploits the structure of the input graphs and specifically, their partitioning properties to parallelize shortest path computations . Our algorithm will be especially efficient if the input graph has a good separator, which means (informally) that it can be divided into two or more equal parts by removing $O(n)$ vertices or edges, where n is the number of the vertices of the graph. Such graphs are frequently seen in road networks, geometric networks and social networks; all planar graphs also satisfy this property. To harness the GPU's parallel computing power for solving the path problem on such graphs, we partition the input graphs into an appropriate number of parts and solve the APSP on each part and then use the partial solutions to compute the distances between all pairs of vertices in the graph. Our algorithm, based on the Floyd-Warshall algorithm, has near quadratic (i.e. near optimal) complexity with respect to the number of nodes, while its matrix-based structure is regular enough to allow for efficient parallel implementation on GPUs. By applying a divide-and-conquer approach, we are able to make use of multi-node GPU clusters, resulting in more than an order of magnitude speedup over fastest known (Dijkstra-based) GPU implementation and a two-fold speedup over a parallel Dijkstra-based CPU implementation.

In what follows, Section II presents recent parallel implementations for solving the APSP problem; in Section III, we detail the principles of our partitioned algorithm; Section IV focuses on the structure of the data and the computations and how the algorithm is implemented on large multi GPU clusters. Finally, Section V shows the results of two experiments and possible ways to improve our implementation.

II. Related Work

When considering a distributed GPU implementation, both the Floyd-Warshall and Dijkstra's approaches have advantages and drawbacks. Though slower for sparse graph, a Floyd-Warshall approach has the advantage of having regular data access patterns that are identical to those of a matrix multiplication. The amount of computations required for a given graph, using a Floyd-Warshall approach, solely depends on the number of vertices in the graph; therefore, balancing workloads between different processing units can be achieved easily. Dijkstra's approach is much faster for sparse graphs but, to achieve best performance, requires complex data structures which are difficult to implement efficiently on a GPU.

Implementing parallel solvers for the APSP problem is an active field of research. Harish and Narayanan [7]

proposed GPU implementations of both the Dijkstra and Floyd-Warshall algorithms to solve the APSP problem and compared them to parallel CPU implementations. Both approaches however require that the whole graph fit in the GPUs memory. They report solving APSP for a 100k vertex graph in around 22 minutes on a single GPU. A cache-efficient parallel, blocked version of the Floyd-Warshall algorithm for solving the APSP problem in GPUs is described in [8]. While the graphs mentioned in [8] are larger than what would fit onto GPU on-board memory, the largest graph instances described in the paper are still only around 10k vertices.

Buluç et al. [9] proposed a blocked-recursive Floyd-Warshall approach. Their implementation, running on a single GPU, shows a speedup of 17-45 when compared to a parallel CPU implementation and outperforms both GPU implementations from [7]. Their blocked-recursive implementation also requires that the entire graph fit in the GPU's global memory; therefore, they only report timings for graphs with up to 8k vertices. Okuyama et al. [10] proposed an improvement over the GPU implementation of Dijkstra for APSP from [7] by caching data in on-chip memory and exhibiting a higher level of parallelism. Their approach showed a speedup of $2.8-13$ over Dijkstra's SSSP-based method of [7]. Matsumoto et al. [11] also proposed a blocked Floyd-Warshall algorithm that they implemented for computations on a single GPU and a multicore CPU simultaneously. Their implementation handles graphs with up to 32k and achieves near peak performance. Only Ortega-Arranz et al. [12] report solving APSP on large graphs - up to 1024k vertices. Using an SSSP-based Dijkstra approach, their implementation runs on a multicore CPU and up to 2 GPUs simultaneously. Experimental work on parallel algorithms for solving just the SSSP problem for large graph instances using a Δ-stepping approach [13] is described in [14].

Our Contribution: We propose a novel APSP algorithm and its parallel implementation to compute all shortest distances between all pairs of vertices of a graph with good partitioning properties. To make the algorithm scalable to large graphs, our implementation uses a combination of shared and distributed-memory GPU computing; the current implementation targets executions on large clusters of GPUs in order to handle graphs with up to a million vertices. Experiments show that the trillion pairs of shortest paths of a million vertex graph can be found in less than 25 minutes using a 64-node cluster with 2 GPUs on each node.

We view our contributions as the following:

(i) We developed a new Floyd-Warshall-based APSP algorithm that is simultaneously work-efficient, has a high-degree of parallelism, and is built upon matrix operations; we are aware of no previous APSP algorithm with such properties.

(ii) Our implementation uses a high degree of parallelism, both at the fine-grained, shared-memory GPU level as well as at the coarse-grained distributed-memory level, employing up to 300 GPUs.

(iii) Our algorithm beats the previous algorithm [12] by more than an order of magnitude with respect to running times using the same or similar computational resources.

(iv) In addition to the fact that our algorithm is faster than Dijkstra-based algorithms, it also has the advantage that it works with arbitrary–negative as well as positive–weights.

III. Algorithm details

In this section we give the overall structure and the idea of the algorithm and describe its individual steps (details of the GPU implementation are covered in Section IV).

A. Overview

Our algorithm takes as input a weighted directed or undirected graph G with n vertices and computes the distances between all pairs of vertices of G. Based on a divide-and-conquer approach, it consists of four steps (see Algorithm 3). In the first step, the original graph G is partitioned into k components of roughly equal sizes using a min-cut like heuristic – our implementation uses a k-way partitioning method from the METIS library [15]. In the second step, the APSP problem is solved on each component independently; in the third step the distance information computed for the components is used to compute distances between all pairs of boundary vertices of G (a *boundary* vertex is one that is adjacent to a vertex from another component); and in the final step the information obtained in steps two and three is combined to compute shortest paths between non-boundary pairs of vertices of G.

We will use the following notation: $\text{dist}_i(v, w)$ will denote the (approximate) value of the distance between v and w computed in Step i, for $i = 2, 3, 4$, and $\text{dist}_G(v, w)$ will denote the (exact) distance in G. Next we describe the steps in more detail.

B. Step 1: Graph decomposition

In Step 1 the input graph G is divided into k components of roughly equal sizes. The decomposition is done be identifying a set of edges (*cut set*) whose removal from G results into a disconnected graph of k parts we call *components*. The set of all components is called a *partition*. Note that while by the standard definition in graph theory, a component is connected, this is not a requirement in our case (although in the typical case our components will be connected). We do require that every vertex in G belong to exactly one component of the partition. Moreover, in order for the resulting APSP algorithm to be efficient, the cut set of edges should be small. Not all classes of graphs

Algorithm 3 Partitioned All-Pairs Shortest Path algorithm

```
INPUT: A graph G(V,E), where V is a set of
    vertices and E a set of weighted edges between
    these vertices.
OUTPUT: The distance of the shortest path between
    any two pairs of vertices in G.

function partitioned_APSP(G)
  // Step 1
  for each Component C in G
      Floyd-Warshall(C) %compute_APSP(C)
  end for

  // Step 2
  Graph BG = extract_boundary_graph(G)
  compute_apsp(BG)
  for each Component C in G
    Floyd-Warshall(C) %compute_APSP(C)
  end for

  // Step 3
  for each Component C1 in G
    for each Component C2 in G
      compute_apsp_between_components(C1, C2)
    end for
  end for
end function
```

have such partitions, but some important classes do. These include the class of planar graphs, the class of graphs of low genus, some geometric graphs, and graphs corresponding to networks with good community structure.

C. Step 2: Computing distances within each graph component

Step 2 involves computing the distances in each component of the partition $\mathcal{P}$ of G using a conventional algorithm, e.g., the Floyd-Warshall or Dijkstra algorithm. For each component $C \in \mathcal{P}$ and any two vertices s and t of C, the output of this step is the minimum length of a path between s and t that is restricted to lie entirely in C. Hence, the distance computed between s and t may be larger that the distances between s and t in G, if there is a shorter path between them that leaves and then re-enters C. Nevertheless, as we will show in later subsections, the computed approximate distances can be used to efficiently compute the correct distances in G.

In order to implement this step, for each component $C \in \mathcal{P}$, a subgraph is extracted containing vertices from the current component and existing edges between these vertices. Any APSP algorithm can then be applied in order to compute distances in each of these sub-graphs. This step thus has k independent tasks–one for each sub-graph–that can be computed in parallel. Since each component contains roughly n/k vertices, using an algorithm whose complexity solely depends on the number of vertices allows these tasks to be computed in roughly the same number of operations. This property can be advantageous depending on the type of parallelism that we want to exploit.

D. Step 3: Computing distances in the boundary graph

In step 3, we first extract the *boundary graph* BG of G with respect to the partition $\mathcal{P}$. The vertices of BG are defined to be all boundary vertices of G. There are two types of edges of BG: the first type are edges in G between boundary vertices from different components. The weights on these edges are the same as their weights in G. The second type of edges, which we call *virtual* edges, are between boundary vertices in the same components – for any two boundary vertices v and w belonging to the same component C there is an edge (v, w) in BG with weight equal to the distance between v and w computed in Step 2. Hence, BG is a compressed version of the original graph, where all non-boundary vertices have been removed, and replaced by edges whose weights are equal to the weight of the shortest path between them. Having constructed BG, we then solve for it the APSP problem using a conventional APSP algorithm.

Despite the fact that the distances encoded in the weights of the new edges of BG are only approximate, the distances between the boundary nodes of BG computed at the end of Step 3 are exact. The next lemma formally establishes this fact.

Lemma 1. *For any two boundary vertices v and w, the distance between v and w in BG is equal to their distance in G.*

Proof: Let $p = (v = x_1, x_2, \ldots, x_l = w)$ be a shortest path in G and let $(x_{b_1}, x_{b_2}, \ldots, x_{b_j})$ be the subsequence of all boundary vertices in p, i.e., $1 = b_1 < \cdots < b_j = l$ and there are no boundary vertices on p between x_{b_i} and $x_{b_{i+1}}$. Hence $p' = (x_{b_1}, x_{b_2}, \ldots, x_{b_j})$ is a path in BG. We are going to estimate the length of p'.

Let $h = (x_{b_i}, x_{b_{i+1}})$ be an edge of p'. If x_{b_i} and $x_{b_{i+1}}$ are from different components, then, by the definition of BG, h is also an edge of G with the same weight as in BG. If x_{b_i} and $x_{b_{i+1}}$ are from the same component C (Figure 1), then h corresponds to a subpath $q = (x_{b_i}, x_{b_i+1}, \ldots, x_{b_{i+1}})$ of p consisting of vertices from only C, by the assumption that p' contains all the boundary vertices of p. Hence, the weight of h and the length of q are the same. By induction on the number of the edges of p', p and p' have the same length, which implies that the distance between v and w in BG is no greater than the distance between them in G. The reverse inequality is obtained in the same way, namely, by showing that any path in BG can be transformed into a path of the same length in G by replacing each virtual edge of the former with the corresponding shortest path computed in Step 2. The claim follows. ■

This step presents no apparent parallelism, since only one task needs to be computed. This absence of parallelism

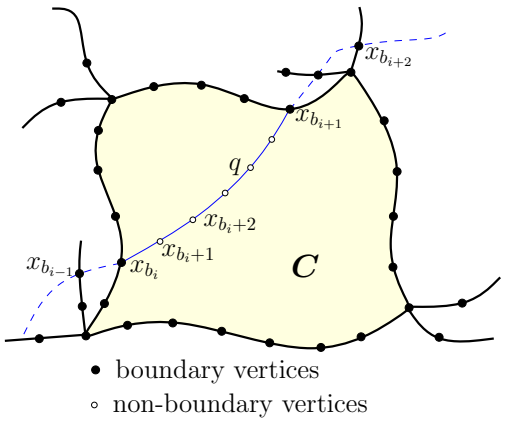

Figure 1. Illustration to the proof of Lemma 1. The shaded region illustrates a component C with the subpath $q = (x_{b_i}, x_{b_i+1}, \ldots, x_{b_{i+1}})$ of p inside it.

at this step may be a major bottleneck for a coarse-grain parallel implementation as boundary graphs can be very large. This issue can however be mitigated by applying our current algorithm recursively on the boundary graph. Boundary graphs are nevertheless denser than the original graph with the addition of virtual edges at Step 2. Boundary graphs are therefore less easily partitioned than input graphs - the number of edges cut per node for a given number of components will be higher.

E. Step 4: Distances between non-boundary vertices

In Step 4 we compute distances where at least one vertex is non-boundary using the information computed in Steps 2 and 3. In order to compute the distance between two non-boundary vertices v_i and v_j from (not necessarily different) components C_i and C_j respectively, we need to find boundary vertices b_i and b_j from components C_i and C_j, respectively, that minimize the sum $\text{dist}_2(v_i, b_i) + \text{dist}_3(b_i, b_j) + \text{dist}_2(b_j, v_j)$, where dist_2 and dist_3 are the distances computed in Step 2 and Step 3, respectively. By our analysis above, dist_3 is the same as the distance in G, but dist_2 is not. We need therefore to prove that such a method produces accurate distances in G.

Lemma 2. *Let v_i and v_j be two vertices from different components C_i and C_j, respectively. Define $B_i = C_i \cap BG$, $B_j = C_j \cap BG$, and*

$$dist_4(v_i, v_j) = \min\{dist_2(v_i, b_i) + dist_3(b_i, b_j) + dist_2(b_j, v_j) \mid b_i \in B_i, b_j \in B_j\}. \quad (1)$$

Then $dist_4(v_i, v_j)$ is equal to the distance in G between v_i and v_j.

Proof: Let p be a shortest path in G between v_i and v_j. Since v_i and v_j belong to different components, then p will contain at least one vertex from B_i and at least one vertex from B_j. Let b_i be the first vertex on p from B_i and b_j be the last vertex on B_j (Figure 2). Let p_1 be the portion of p between v_i and b_i, p_2 be the portion between b_i and b_j, and p_3 – the portion between b_j and v_j. Since any subpath of a shortest path is also a shortest path between the corresponding endpoints, p_1 is a shortest path in G between v_i and b_i, i.e., $|p_1| = \text{dist}_G(v_i, b_i)$. Moreover, by the definition of b_i as the first boundary point of C_i on p, p_1 is entirely in C_i and hence $|p_1| = \text{dist}_2(v_i, b_i)$. In the same way one can prove that $|p_2| = \text{dist}_2(b_j, v_j)$. Finally, $|p_3| = \text{dist}_G(b_i, b_j) = \text{dist}_3(b_i, b_j)$ by Lemma 1. Hence

$$|p| = |p_1| + |p_2| + |p_3| = \text{dist}_2(v_i, b_i) + \text{dist}_3(b_i, b_j) + \text{dist}_2(b_j, v_j).$$

By the definition of $\text{dist}_4(v_i, v_j)$ as a minimum over all $b_i \in B_i, b_j \in B_j$, the last equality implies $\text{dist}_4(v_i, v_j) \leq \text{dist}_G(v_i, v_j)$. But since $\text{dist}_4(v_i, v_j)$ is a length of a path between v_i and v_j, while $\text{dist}_G(v_i, v_j)$ is the length of a shortest path, then $\text{dist}_4(v_i, v_j) \geq \text{dist}_G(v_i, v_j)$. Combining the last two inequalities we infer that none of them can be a strict inequality, i.e., $\text{dist}_4(v_i, v_j) = \text{dist}_G(v_i, v_j)$. ■

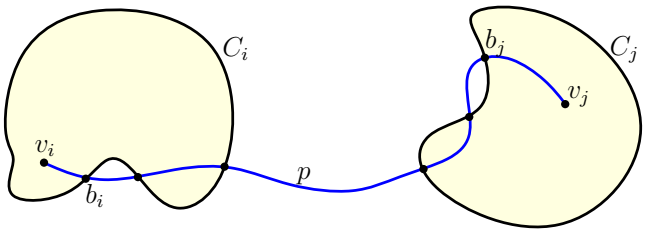

Figure 2. Illustration to the proof of Lemma 2. Note that while in the figure both v_i and v_j are non-boundary, the proof does not make such an assumption.

Lemma 3. *Let v_i and v_j be two vertices from component C_i. Then $dist_G(v_i, v_j) = \min\{dist_2(v_i, v_j), dist_4(v_i, v_j)\}$, where $dist_4$ is as defined in Lemma 2.*

Proof: Consider the following two cases. If p leaves C_i, then p should cross the boundary B_i at least twice. Define b_i and b_j as the first and last vertex from B_i on p. Then exactly the same arguments as in Lemma 2 apply to the three paths into which b_i and b_j divide p. In this case $\text{dist}_G(v_i, v_j) = \text{dist}_4(v_i, v_j)$. If p does not leave p, then Step 2 will compute the accurate distance in G between v_i and v_j, and therefore $\text{dist}_G(v_i, v_j) = \text{dist}_2(v_i, v_j)$. ■

The lemmas imply that the distances in G between all pairs of vertices where at least one of the vertices is non-boundary can be computed by using (eq:step4). Since we don't know which pair (b_i, b_j) of boundary nodes corresponds to the minimum in (eq:step4), we have to try all such pairs, resulting in total of $|B_i||B_j|$ operations needed for computing $\text{dist}_G(v_i, v_j)$. For a graph with k components, we need to compute the distances between pairs in any pair of components; we therefore have k^2 independent tasks. Components being of roughly equal sizes, these tasks also represent the same amount of computations. This step is

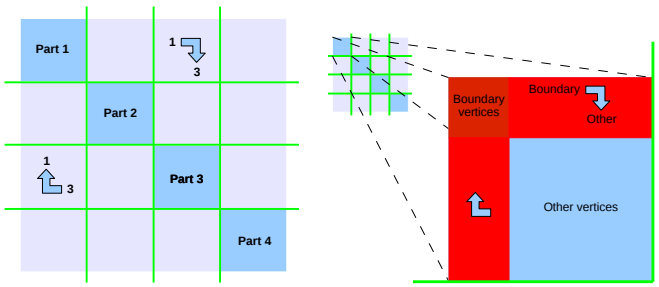

Figure 3. Adjacency matrix after reordering of the vertices. Vertices from the same component are stored contiguously starting with boundary vertices (in red).

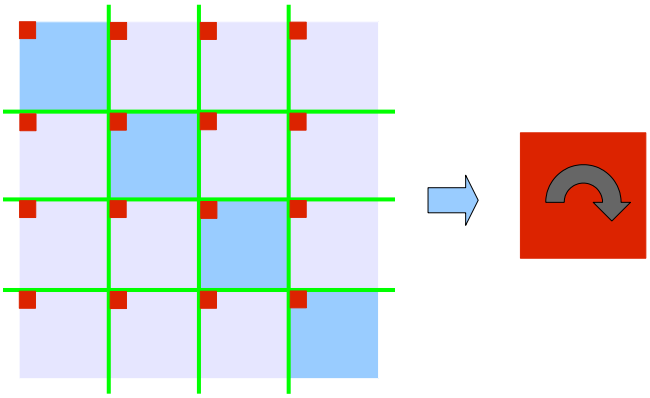

Figure 4. The boundary matrix, here in red, is scattered over the adjacency matrix. Step 3 consits in reconstituting the boundary matrix and computing shortest distances.

the most computationally intensive, but presents massive, already balanced, coarse-grain parallelism.

IV. Implementation

In this section, we first focus on how operations described in the previous section translate in terms of data structures. We then detail the two-level parallel aspect of our implementation. We finally describe the current main memory bottleneck of our approach.

A. Data organization

A simple way to represent a weighted graph is to use an adjacency matrix. For very large graphs however, such a memory intensive representation is often avoided. Instead, large sparse graphs are stored using lists; sub-matrices, corresponding to sub-graphs, are extracted from these lists. For simplicity reasons, we can however assume that a large adjacency matrix representation is available and keep in mind that sub-matrix extraction operations are slightly more costly than they appear. We are also taking into account the fact that, even when the input graph (matrix) is sparse, the output is always a dense matrix as it encodes the distances between all pairs of vertices.

Partitioning the graph is performed using a k-way partitioning routine from the METIS library [15]. The result is a partitioning of the graph into k parts such that the number of edges with endpoints in different parts is minimized. Since that partitioning problem is NP-hard, METIS computes an approximation based on heuristics. Vertices are then reordered so that vertices belonging to the same component are numbered consecutively starting with the boundary vertices – see Figure 3.

Diagonal sub-matrices contain information about sub-graphs for each component; non-diagonal sub-matrices contain known shortest distances between components. Within each diagonal sub-matrix, the top left sub-matrix contains information about the sub-graph induces by boundary vertices of the component; the bottom right sub_matrix contains information about the sub-graph induced by non-boundary vertices of the component and the rest of the diagonal sub-matrix contains known shortest distances between boundary and non-boundary vertices.

For Step 2, diagonal sub-matrices are extracted; a Floyd-Warshall approach is then used to compute shortest distances. The Floyd-Warshall algorithm guarantees that the total number of operations for a single matrix solely depends on the size of the matrix. Since all components of the graph have roughly the same number of vertices, all diagonal sub-matrices represent roughly the same amount of operations.

For Step 3, the boundary matrix is extracted – see Figure 4. We then apply the same algorithm recursively reducing the number k of component at each iteration. Recursion stops when $k = 1$ or when the boundary graph becomes so dense that it does not have good partitioning (in terms of number of boundary vertices). At that point the APSP subproblem is solved using Floyd-Warshall.

For Step 4, we compute shortest distances between every pair of distinct components. This process corresponds to filling non-diagonal sub-matrices. For two components I and J, filling the associated, I to J, non-diagonal sub-matrix requires information from three sub-matrices:

- the non-diagonal sub-matrix being filled. We are particularly interested in the part of the sub-matrix containing shortest distances between boundary vertices from component I to boundary vertices from component J.
- the diagonal sub-matrix corresponding to component I - located in the same row as the non-diagonal sub-matrix being filled. We are particularly interested in the part of this diagonal sub-matrix that contains shortest distances from any vertex of component I to boundary vertices.
- the diagonal sub-matrix corresponding to component J - located in the same column as the non-diagonal sub-matrix being filled. We are particularly interested in the part of this diagonal sub-matrix that contains shortest

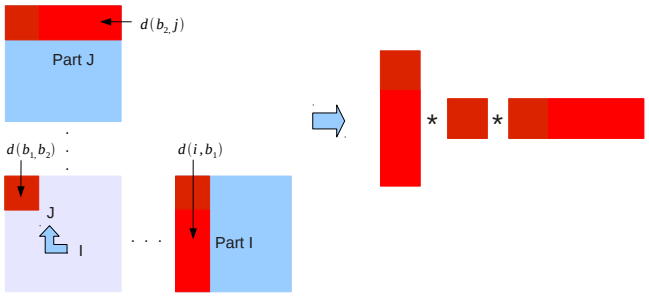

Figure 5. Computations associated to each non-diagonal sub-matrix uses data from 2 diagonal sub-matrices and part of the non-diagonal sub-matrix itself. Computations are similar to matrix multiplications.

distances from boundary vertex of component J to any vertex - see left of Figure 5.

Shortest distances from vertices from component I to vertices from component J are obtained by multiplying the three parts of sub-matrices - as shown on the right of Figure 5 - where $(+, *)$ operations are replaced with $(min, +)$ operations.

B. Work analysis

Next we will try to estimate the work (number of operations) of the algorithm. Since the work depends on the partitioning properties of the input graph, we will do the analysis for the case of planar bounded-degree graphs. For that class of graphs, there exists a partitioning of any n-vertex graph into k parts such that the number of boundary vertices in each part is $O(\sqrt{n/k})$ [16]. We make the assumption that METIS produces a partition with such properties. Although the partition METIS produces does not come with theoretically guaranteed bounds, it works in practice better than alternative algorithms that have such guarantees, which is the reason we chose it. The time needed for Step 1 is $O(n \log n)$.

In Step 2, we have k subtasks of computing APSP on graphs of size $O(n/k)$ using an algorithm of cubic complexity, so the number of operations for that step is $k(n/k)^3 = n^3/k^2$.

In Step 3, we have to solve the APSP on a graph of size $O(k\sqrt{n/k}) = O(\sqrt{kn})$. Using an algorithm with complexity $O(N^\alpha)$, where N is the number of the vertices of the subgraph, the number of operations for this step is $O((kn)^{\alpha/2})$.

For Step 4, we have k^2 tasks and each tasks involves the multiplication of three matrices with dimensions $n/k \times \sqrt{n/k}$, $\sqrt{n/k} \times \sqrt{n/k}$, and $\sqrt{n/k} \times n/k$, respectively. Computing the product of the first and the second matrix takes

$$O((n/k)\sqrt{n/k}\sqrt{n/k}) = O((n/k)^2)$$

operations and finding the product of the resulting $n/k \times \sqrt{n/k}$ matrix and the third matrix takes

$$O((n/k)\sqrt{n/k}(n/k)) = O((n/k)^{5/2})$$

operations, which is the dominating term. Hence, the total number of operations for Step 4 is

$$O(k^2(n/k)^{5/2}) = O(n^{5/2}/k^{1/2}).$$

The total number of operations is the sum of the numbers computed for Steps 1, 2, 3, and 4 and is minimized when $(kn)^{\alpha/2} = n^{5/2}/k^{1/2}$ or $k^{\alpha+1} = n^{5-\alpha}$. If in Step 3 Floyd-Warshall is used, then $\alpha = 3$ and $k = n^{1/2}$ is optimal, resulting in a bound of $O(n^{9/4})$ for the total number of operations, slightly worse than the theoretical lower bound of $O(n^2)$. Our implementation in fact uses recursion in Step 3 so the total complexity is even closer to quadratic, but we will skip the details of the exact evaluation since the analysis gets much more complex.

C. Parallel implementation

Our implementation specifically targets large clusters of hybrid systems - possessing both a multicore CPU and manycore GPUs. This implementation exploits parallelism at two levels. At a coarse-grain level, large independent tasks - corresponding to computations of diagonal and non-daigonal sub-matrices - can be performed simultaneously on different nodes of a cluster. At a fine-grain level, each task is computed on a massively parallel GPU. Remaining CPU cores handle tasks that are not suited for GPUs: input/output file operations and communication with other nodes.

Coarse-grain parallelism: Steps 2 and 4 of our algorithm exhibit interesting parallel properties: a large number of balanced, independent tasks; k tasks for Step 2 and $k^2 - k$ for Step 4. Using the MPI standard [17], these tasks are distributed accross nodes of the cluster for simultaneous computations. One master node is in charge of reading the input graph file, calling the partitioning routine and sending tasks to a number of slave nodes equal to the number of available GPUs on the cluster. Depending on the cluster's topology, the number of master and slave nodes will not match the number of physical nodes used on the cluster if each cluster node contains more than one GPU.

For Step 3, the large initial boundary matrix is computed recursively using the same algorithm with decreasing values for the number k of components. The amount of independent tasks therefore decreases with k, until a single, smaller boundary matrix is obtained and computed by a single slave node.

Fine grain parallelism: Upon receiving a task from the master node, each slave node then sends the corresponding data to its GPU for computations, retrieves results and send them back to the master node. Tasks are of two different kinds: diagonal workloads, which consist in computing shortest distances over a small subgraph, and non-diagonal workloads, which consist in multiplying three matrices.

Computations of diagonal workloads are implemented on the GPU using a blocked-recursive Floyd-Warshall approach developped by [9] and adapted for non-power of 2 matrices. Non-diagonal workloads require less synchronization and can be implemented using a fast matrix-multiplication approach derived from [18] and adapted for $(min, +)$ operations.

In this configuration, each physical node on the cluster makes use of as many CPU cores as there are available GPUs. If more CPU cores are available than GPUs, computational power is still available. On slave nodes, remaining CPU cores are used for outputting final results to disk. On large clusters, communication between the master node and slave nodes can become a bottleneck, leaving slave nodes idle while waiting for the master node to be available. In order to increase the availability of the master node, a single CPU thread is used to initiate communications with slave nodes while remaining CPU cores handle the rest of the communications, updating data structures with temporary results and outputting final results to disk.

D. Memory limitations

For very large input graphs, memory usage becomes an issue. As stated previously, an entire adjacency matrix for the graph cannot be allocated; the graph is instead kept in memory as a list of edges, a much more memory-efficient representation. Even with this efficient representation, temporary sub-matrices (diagonal sub-matrices and boundary matrices) need to be kept in memory. When recursively computing Step 3, boundary matrices are output to files so as to only keep a single boundary matrix in memory.

Final results for diagonal sub-matrices are only obtained at the end of Step 3. As soon as final values for these diagonal sub-matrices are obtained, they are output to files; only relevant parts are kept in memory for Step 4, namely, parts of these sub-matrices containing shortest distances from and to boundary vertices. Shortest distances between non-boundary vertices are thus discarded from main memory at the end of Step 3.

The current limiting factor in terms of memory usage is the initial boundary matrix. The first boundary matrix has to fit in the main CPU memory. Section V discusses ways to overcome this limitation. It is however probable that prohibitive run-times or results too large to process may become the limiting factor before main memory usage does.

V. Results and perspectives

In this section, we compare our implementation to two parallel Dijkstra implementations. It is important to note that our implementation allows graphs with negative edges – but no negative cycles – unlike Dijkstra-based approaches.

In order to test our implementation, we generated random graphs with increasing numbers of vertices, ranging from 1024 to 1024k. These graphs, generated using the LEDA library [19], were made planar to ensure good partitioning properties.

Figure 6. Evolution of run times with respect to the number of vertices. Two implementations are compared: our implementation using external memory and the GPU Dijkstra implementation from [12]. Computations were run using two GPUs on a single cluster node.

Computations were run on a cluster of more than 300 computer nodes; each node is equipped with two NVIDIA C2090 GPUs, a 16 core Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz and 32 GB of RAM.

Our implementation handles instances up to 512k vertices without using external memory. For the very last instance, the use of external memory was required to fit in the 32 GB of main memory. We later refer to our implementation without using external memory as "Part. APSP no EM" and our implementation using external memory as "Part. APSP EM".

The GPU Dijkstra implementation from [12] is, to the best of our knowledge, the only implementation that was reported to solve APSP for graphs with up to 1024k vertices; we later refer to this implementation as "GPU Dijkstra". This implementation parallelizes SSSP computations on a single computer using two GPUs and a multicore CPU. In order to compare this implementation to ours, we restricted computations of both implementations to using only two GPUs. Both implementations could therefore run on a single cluster node; no communication between nodes were therefore required.

Figure 6 shows the runtimes for GPU Dijkstra and Part. APSP EM for graphs with numbers of vertices ranging from 1024 to 1024k using only two GPUs. GPU Dikstra could not compute the last two instances - 512k and 1024k vertices - within the 10 hour limit enforced on the cluster. Results in 6 indicate our implementation to be significantly faster than GPU Dijkstra.

Figure 7 shows the evolution of the speedup of our method without using external memory with respect to the number of GPUs used for the computations. Speedups are calculated using the run time obtained using only one GPU as a reference. Computations were done for the 512k vertex instance using the Part. APSP (without I/O). We can see

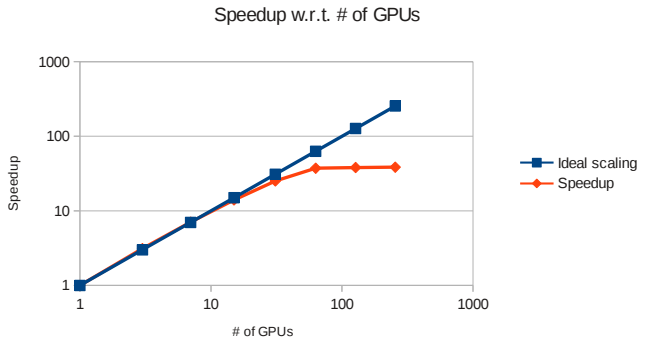

Figure 7. Evolution of speedups with respect to the number of GPUs. The ideal scaling line is given as a reference.

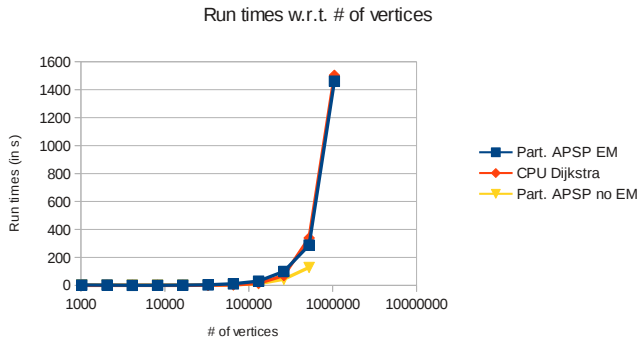

Figure 8. Evolution of run times with respect to the number of vertices. Three implementations are compared: our two implementations - with and without using external memory - and a distributed Dijkstra implementation referred to as CPU Dijkstra. All computations were run on 64 cluster nodes.

that coarse-grained parallelism is close to optimal up to around 31 GPUs; almost no benefit can however be gained from using more than about 63 GPUs. The reason for this stagnation of the speedup above 63 GPUs is the saturation of communication with the master node.

The scalability can be improved using a coarse-grained parallelism approach that would relieve the master node of some of its communication. A work-stealing approach, for instance, would reduce the amount of communication required for the master node by decentralizing some of the memory transfers. A work-stealing approach is however difficult to implement, due to the two-sided communication scheme enforced by the MPI standard. [20] showed that such an efficient approach was nevertheless feasible. This issue could also be addressed by creating a hierarchy of master nodes; some computations would be redundant between the different master nodes - handling the main data structure - but this would only represent a negligible fraction of the overall workload.

Figure 8 shows a comparison between our two implementations and a distributed Dijkstra approach - later referred to as CPU Dijkstra - for graphs ranging from 1024 to 1024k vertices. The distributed Dijkstra approach was implemented by dynamically distributing SSSP computations for each vertex of the graph over every core of every available cluster node. The Dijkstra-based implementation used is that of the Boost C++ library [21]. This experiment is not intended to compare directly the performances of 2 GPUs versus a multicore CPU. Instead, we intend to show that our approach is competitive with a distributed Dijkstra approach given a fixed number of heterogeneous cluster nodes. The run times presented in Figure 8 were obtained using 64 cluster nodes. We can see that our version using external memory obtains very similar run times to that of the distributed Dijkstra version, while allowing graphs with negative edges to be computed. Our version without external memory is however significantly faster.

In order to test our implementation on a real dataset, we retrieved the Californian road network dataset from [22]. This dataset consists in the entire road network of the state of California; it contains $1,957,027$ vertices corresponding to road intersections and more than 5 million edges corresponding to roads. Computing the 3.8 trillion shortest distances in this network took 31 minutes, using 64 cluster nodes.

VI. Conclusion

We described a new algorithm for solving the all-pairs shortest path problem on planar and other graphs with good partitioning properties, which is characterized by a nearly optimal number of operations, a regular matrix-structured computations, and which admits a high degree of parallelism. Our implementation on a multi-GPU cluster allows a trillion distances to be computed in half an hour or less. Compared with similar algorithms, ours is orders of magnitude faster and also allows exploiting a much larger number of GPUs. Our future work will target improving the coarse-grained communication structure and increasing the memory efficiency so that even larger instances can be run without using external memory.

References

[1] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[2] T. H. Cormen, C. Stein, R. L. Rivest, and C. E. Leiserson, *Introduction to Algorithms*, 2nd ed. McGraw-Hill Higher Education, 2001.

[3] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.

[4] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 641–650.

[5] V. Traag and J. Bruggeman, "Community detection in networks with positive and negative links," *Physical Review E*, vol. 80, no. 3, p. 036115, 2009.

[6] K. Inoue, A. Doncescu, and H. Nabeshima, "Hypothesizing about causal networks with positive and negative effects by meta-level abduction," in *Inductive Logic Programming*. Springer, 2011, pp. 114–129.

[7] P. Harish and P. Narayanan, "Accelerating large graph algorithms on the GPU using CUDA," in *High performance computing–HiPC 2007*. Springer, 2007, pp. 197–208.

[8] G. J. Katz and J. T. Kider, Jr, "All-pairs shortest-paths for large graphs on the gpu," in *Proceedings of the 23rd ACM SIGGRAPH/EUROGRAPHICS symposium on Graphics hardware*, ser. GH '08. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2008, pp. 47–55.

[9] A. Buluç, J. R. Gilbert, and C. Budak, "Solving path problems on the GPU," *Parallel Computing*, vol. 36, no. 5, pp. 241–253, 2010.

[10] T. Okuyama, F. Ino, and K. Hagihara, "A task parallel algorithm for finding all–pairs shortest paths using the GPU," *International Journal of High Performance Computing and Networking*, vol. 7, no. 2, pp. 87–98, 2012.

[11] K. Matsumoto, N. Nakasato, and S. G. Sedukhin, "Blocked united algorithm for the all-pairs shortest paths problem on hybrid CPU-GPU systems," *IEICE TRANSACTIONS on Information and Systems*, vol. 95, no. 12, pp. 2759–2768, 2012.

[12] H. Ortega-Arranz, Y. Torres, D. R. Llanos, and A. Gonzalez-Escribano, "The all-pair shortest-path problem in shared-memory heterogeneous systems," 2013.

[13] U. Meyer and P. Sanders, "Delta-stepping: a parallelizable shortest path algorithm." *J. Algorithms*, vol. 49, no. 1, pp. 114–152, 2003. [Online]. Available: http://dblp.uni-trier.de/db/journals/jal/jal49.html#MeyerS03

[14] K. Madduri, D. A. Bader, J. W. Berry, and J. R. Crobak, "An experimental study of a parallel shortest path algorithm for solving large-scale graph instances." in *ALENEX*. SIAM, 2007. [Online]. Available: http://dblp.uni-trier.de/db/conf/alenex/alenex2007.html#MadduriBBC07

[15] G. Karypis and V. Kumar, "Multilevel k-way partitioning scheme for irregular graphs," *Journal of Parallel and Distributed computing*, vol. 48, no. 1, pp. 96–129, 1998.

[16] G. N. Frederickson, "Fast algorithms for shortest paths in planar graphs, with applications." *SIAM J. Comput.*, vol. 16, no. 6, pp. 1004–1022, 1987.

[17] M. Snir, S. W. Otto, D. W. Walker, J. Dongarra, and S. Huss-Lederman, *MPI: the complete reference*. MIT press, 1995.

[18] V. Volkov, "Better performance at lower occupancy," in *Proceedings of the GPU Technology Conference, GTC*, vol. 10, 2010.

[19] K. Mehlhorn, S. Näher, and C. Uhrig, "Leda: A platform for combinatorial and geometric computing," vol. 38, 1999.

[20] G. P. Pezzi, M. C. Cera, E. Mathias, and N. Maillard, "On-line scheduling of MPI-2 programs with hierarchical work stealing," in *Computer Architecture and High Performance Computing, 2007. SBAC-PAD 2007. 19th International Symposium on*. IEEE, 2007, pp. 247–254.

[21] B. Dawes, D. Abrahams, and R. Rivera, "Boost C++ libraries," *URL http://www. boost. org*, vol. 35, p. 36, 2009.

[22] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.

VII. Acknowledgments

The authors acknowledge and appreciate the support provided for this work by the Los Alamos National Laboratory Directed Research and Development Program (LDRD). This work was also partially supported by the region of Brittany, France.