# Understanding the Performance of Parallel Graph Applications in a Tiered-memory System

**Milestone Report**
**Kaiwen Xue. Andrew ID: kaiwenx**

## 1. Updated Schedule

The schedule in my proposal is updated in Table 1. I have completed all items listed in the proposals due before today, and am in the process of completing items due this week.

| Date | Area | State |
|---|---|---|
| Wed Nov. 22 | Hardware simulation environment | Done |
| Wed Nov. 29 | Motivation study | Done |
| Wed Dec. 6 | Multicore characterization | In progress |
| Thu Dec. 14 (Deadline) | Algorithm design | In progress |
| Thu Dec. 14 (Deadline) | Report writing | Not started |

Table 1: Updated schedule

## 2. Current Results

### 2.1. Completed Items

I have built a hardware simulation environment using modified QEMU, SST, and Linux kernel. Currently, I am able to perform full system simulation by running the kernel with QEMU, which sends memory instructions to the SST for cycle-accurate simulation. The environment can simulate the parallel operation of up to 8 cores. I plan to increase the maximum number of supported cores if needed. With this framework, any proposed hardware changes and instrumentation can be implemented in SST, and any proposed software changes can be implemented in the guest kernel or application.

### 2.2. Findings

Using the framework, I have performed studies of the behavior of two applications on their memory access patterns. Detailed results will be demonstrated in the poster session and the final report, as they will take up significant space for a discussion of satisfactory quality. In general, I have found the following:

- It is possible to propose architectural changes from the MMU (TLB or page table cache) to provide insights to the hotness of a particular region of memory, and it is possible to be done per-core
- TLB misses of all cores will have more severe effect when a tiered-memory system is employed
- Hotness within a page may be very imbalanced, both per-core and globally

### 2.3. Challenges

Existing literature has proposed interesting ideas on other ways of measuring access hotness and using that information to improve application performance. A challenge is to compare and contrast the advantages and limitations of these approaches. A section of this discussion will be included in the final report and the poster session, if meaningful results can be gained before the final report deadline. Otherwise, this discussion will be left as future work.

## 3. Future Work

### 3.1. Work Plan

I will focus on studying how a multicore system interplays with a tiered-memory system. Will a core be pinged on data that is in the CXL (remote) memory? Will the work become more imbalanced if a core is using inefficient memory, and how will it effect the end-to-end performance?

### 3.2. Plan for Poster Session

I plan to share with the instructors and classmates the following during the poster session. Each section will take up 2 pages on the poster:

1. Design of the simulation infrastructure
2. Selected important results of my motivation studies mentioned in Section 2.2
3. Results of the multicore studies
4. Proposal and demo of the algoritm for smart page placement. I expect it to be ongoing before the semester ends, but a first iteration of the algorithm should be completed by then, even if the performance is unsatisfactory.