The government seeks individual input; attendees/participants may provide individual advice only.

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes
July 12, 2023

Virtual Meeting

Participants

Alejandro Suarez (NSF)	Jay Park (NSF)
Andrey Kanaev (NSF)	Jonathan Skone (NERSC)
Arjun Shankar (ORNL)	Mallory Hinks (NCO)
Brad Settlemyer (NVIDIA)	Marcy Collinson (Oracle)
David Martin (ANL)	Marlon Pierce (NSF)
Dhruva Chakravorty (Texas A&M)	Michael Corn (NSF)
Hal Finkel (DOE)	Mishra Sambit
Ian Karlin	Val Anantharaj (ORNL)
Jack Wells (NVIDIA)	

Introductions: This meeting was chaired by Jay Park (NSF) and Hal Finkel (DOE SC)

Towards Data Centric Discovery with High Performance Computing
Brad Settlemyer, NVIDIA

Brad discussed the evolution of High-Performance Computing (HPC) and Cloud Service Providers (CSPs) platforms, moving towards larger, more capable nodes with more cores. NVIDIA is developing a DPU accelerated server where jobs are run on hosts, and infrastructure and other work that do not require an actual core are moved onto some other type of core. The DPU is used to accelerate collectives and storage systems to push compression off.

He also outlined five principles to programming DPUs. Firstly, right-sizing the work performed on a DPU. Secondly, treating the DPU like a first-class peer when possible, which means communicating directly with the DPU rather than delegating tasks. Thirdly, leveraging advanced DPU features, such as diverting data directly into host memory or implementing line-rate compression. Fourthly, using the DPU to provide useful services that can make a CPU or GPU go faster, such as scheduling and load balancing. And finally, revisiting old assumptions about specialization and abstraction to pry open the abstractions and get data out to a DPU or somewhere in the data so that DPU-based processes can assist.

Brad highlighted the need to build a data analysis platform that maps scientific data onto a standard workflow and commits to building a platform for doing that type of analysis. It also emphasizes the importance of adopting as much of what already exists as possible and not trying to rebuild everything. The next significant improvement in data analytics is predicted to come from near-processing storage, which will allow for more flexible pushing of predicates along the data path rather than just into near storage.

Questions:

- David Martin asked Brad for his thoughts on how DAOS plays with the sort of architecture he described.
    - Brad said DAOS doesn't have a path for fast read analytics, as it stands today. So, it is a successor primarily to fixing some of the limitations that existed for Lustre. And a lot of those limitations have turned out not to be limitations. However, what is great about DAOS is that it does have these side channel paths to accessing the storage. One, you can map tasks onto the DAIS object servers. It has the notion of extremes, which are these execution streams that can execute at the storage server. And two, it also has the notion of objects having a location.
    - He said at present, it doesn't seem to me to have the features that would allow it to do latency-optimized storage.
- Dhruva Chakravorty said he had a question about the in-network computing. It sounded really exciting. He asked if Brad could talk a little bit about the limitations of the PCIe switch and the generation of PCIe and how it impacts the offloading capabilities here? He said he's curious about how that impacts speeds.
    - Brad said this is interesting. The problem is that if you have a box that's built with NVIDIA Bluefields, they are what are called target-mode Bluefields, and they actually exist without having another CPU in the box. Whenever you try and build a store-reliable box of flash, you end up having to have two PCIe switches so that you can be sure you have failover.
    - One of the remaining challenges is figuring out how to right-size, how much PCIe belongs inside the box as you try an impedance match, the network, the CPUs, which putting the CPUs along the data path at least prevents you from having to have an IOH that is also going to take more lanes out of your PCIe switch.
    - Now for NVMe SSDs, we kind of know the next generation of SSDs are going to look like, but if you were to put a bunch of GPUs and more things like that in here, you do have to think about the richness of the PCIe fabric that you need inside the node.
    - In particular as nodes get larger, that becomes one of the system design, one of the most important system design parameters.
- Dhruva asked if Brad saw CXL changing this in the near future of if he thought this would still be PCIe driven?
    - Brad said he like this idea of the ability to take a memory buffer and share it amongst all the things that exist inside of this box. He thinks that's a great place for his data analytics point. He said if he were building a large distributed index that described a lot of data, he'd probably want to land that in memory and be able to work on it and then land it back into storage.
    - Brad said he thinks there's big capabilities for using something like CXL to have a smart scratch pad for data analytics and doing index processing and bitmap joining and a lot of the steps that you need to do for very fast analysis.
- Val Anandaraj had a question from a higher level perspective as an end user who's doing some data intensive computing. He said just last week, he reduced about 200 terabytes of climate data to crunch out, just come up with just 10 figures just to do some validation studies and he used 82 nodes on a generic Linux cluster and the job got done in about 75 minutes. It was a good day for the file system, so the file system did most of the heavy lifting. The computing wasn't really the bottleneck, but it reduced that kind of time by an order of magnitude, for example. He said that Brad had mentioned Parquet in one of his slides, and asked what are the paradigms that he can think about for the science users? A lot of them use HDF5, a lot of them chunk and group the data

for some kind of to optimize the IO, either to write it or to read it. He asked Brad if there was another level that he could talk about for scientific use cases?

- Brad said that he thinks that the question is rooted in the theory that Val wants to analyze the data that he generates. Brad said that if that climate data has longstanding data and that other people would have liked to have analyzed it, rather than just dump out the 10 numbers what he wishes would have happened in his ideal world is probably impossible to do this in this system.
- He said what would be the ideal is Val then would have dumped out a set of Parquet data that was in Presto or Impala format, and then anybody could run that query.

Composabilty Across Dis-aggregated Infrastructure NSF FASTER and ACES
Dhruva Chakravorty, TAMU

Dhruva discussed the performance of composable HPC, in comparison to traditional HPC layouts. He introduced FASTER (Fostering Accelerated Scientific Transformations, Education, and Research) and ACES (Accelerating Computing for Emerging Sciences). FASTER was put through various benchmarks, to gauge its performance scaling. The results indicated that it was able to provide good performance scaling, proving to be a strong contender against traditional HPC clusters.

However, some performance drops were observed, and further investigation is needed to understand the cause. The study also highlighted the need for new benchmarks to test and configure composable machines, as the traditional approaches may need to be revisited. The researchers also emphasized the need for mechanisms to identify the best composed hardware choices, as this is crucial for researchers to optimize their workloads.

Dhruva concluded with a call to action for researchers to use the machines and apply for allocations and access, as well as an invitation to attend the PEARC23 conference, where the researchers will be presenting six papers, two tutorials, and a BOF session. He also expresses gratitude towards the National Science Foundation for funding ACES, FASTER, and Sweeter and acknowledges the incredible team working on the project.

Questions:

- Jonathan Skone asked about the underlying physical architecture of the nodes. He asked if they were all homogeneous. He also asked about the GPU-CPU ratio and the network.
    - Dhruva said FASTER has 180 nodes, HS256 gigs of RAM, about 11,000 cores. Each compute node is designed to see on processors. He said he can get the networking layout to him off line.
- Jay Park asked about the configuration of the liquid architecture and its affect on the performance of the scalability of the GPU.
    - Dhruva said the scalability depends largely on the workload, the amount of data. He said when they run HPL, they get pretty good scale on liquid. HPL runs well, stream runs well, when they start going to these research workloads, as long as they are in AI/ML space, mostly everything works well. But when they go into codes like PyFR, things start breaking and they aren't sure why.

Roundtable

- Marcy Collinson gave an update on Oracle's sponsorship with the Research Data Alliance. She shared a link to an update on the first working group, which is centered around mapping the landscape of digital research tools. https://www.rd-alliance.org/kickstarting-rda-ofr-working-group-map-landscape-digital-research-tools She said if anyone has interest in that, please let her know. She said they are getting ready to launch their second working group with them. https://www.rd-alliance.org/new-rda-working-group-managing-and-integrating-multiomics-data She said hat will be centered around identification of multiomics data standards. She also said they are going to be launching the workshop soon and would love to see some MAGIC members in attendance.
- David Martin asked about a potential MAGIC meeting at SC23 in Denver.
    - Mallory said she believes we will be holding an in-person meeting there again this year.
- Jay Park thanked Mallory and MAGIC attendees for their help and interest during his tenure as MAGIC co-chair, which will be over at the end of July.

Next Meeting August 2, 2023