TECHNICAL OVERVIEW

MAXIMIZING DATA CENTER PRODUCTIVITY WITH APPLICATION WORKLOAD ANALYSIS

Introduction

High performance computing (HPC) is a fundamental pillar of modern science. From predicting weather to discovering drugs to finding new energy sources, researchers use large computing systems to simulate and predict our world. Improving the productivity and increasing the number of scientific simulation iterations can have a profound impact on the quantity and quality of breakthroughs. Recent examples of scientific breakthroughs at supercomputing sites include modeling the formation of the HIV capsid, predicting weather, and accelerating life-saving drug discoveries.

Detailed knowledge of application workload characteristics can optimize performance of current and future systems. This may sound daunting, with many HPC data centers hosting over 2,000 users running thousands of applications and millions of jobs per month. However, at key supercomputing sites, a common pattern has emerged. Less than 2 percent of applications occupy most of the time on the system. This makes it relatively easy to understand the benefit of accelerated computing. In short, a small amount of analysis can yield millions of dollars in savings or the ability to buy a much more powerful and capable supercomputing system.

Performing Application Workload Analysis

At a high level, the process of profiling a data center's application workload to maximize data center throughput includes three steps:

Step 1

Measure System Application Profile

Profile workloads by application for a time period long enough to represent the overall user base

Step 2

Identify Key Apps

Identify the applications that consume most of the available system.

Step 3

Explore New Applications

Assess new approaches such as data science, machine learning (ML), artificial intelligence (AI), and shifts in historical application allocations.

Step 1: Measuring Current System Profile

Application workload analysis involves more than just knowing which scientific domains or applications are running on the system. It requires a detailed understanding of application-level system usage and throughput. Data center throughput is measured as the number of scientific jobs that can be achieved per day. This is different than the traditional way of measuring throughput as the number of floating point operations per second (FLOPS) generated using synthetic

benchmarks like the linear equation software package (Linpack) or high-performance conjugate gradients (HPCG).

Supercomputing sites such as the National Energy Research Scientific Computing Center (NERSC), the Texas Advanced Computing Center (TACC), and the National Center for Supercomputing Applications (NCSA) have performed application workload analyses to understand their current workload profiles and anticipate the future needs of their users. When designing a new system, it's important to identify processes and tools that can accurately collect and analyze application usage data without impacting system performance. For example, TACC has developed a tool called XALT that allows supercomputer support staff to collect and understand job-level information about the libraries and executables that end users access. XALT collects accurate, detailed data and stores that data in a database; all the data collection is transparent to the users. According to the *2017 Workload Analysis of Blue Waters Report*, there were over 35,000 node hours consumed on the Blue Waters supercomputer to analyze roughly 95 terabytes (TB) of input data from more than 4.5 million jobs over a 3.5-year period. A workflow pipeline was established so that data from all future Blue Waters jobs will be automatically ingested into the Open XDMoD data warehouse, making future analyses much easier.

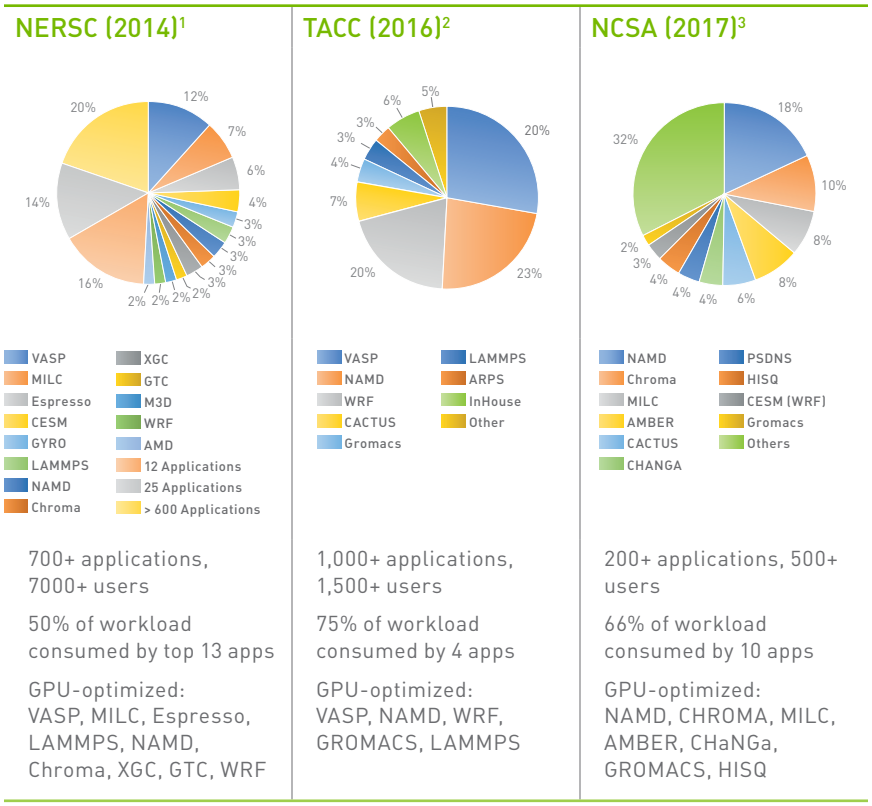

Figure-1: System utilization at three major sites for applications by system hours

1. Source: http://portal.nersc.gov/project/mpccc/baustin/NERSC_2014_Workload_Analysis_v1.1.pdf

2. Source: http://web.corral.tacc.utexas.edu/XALT/

3. Source: https://arxiv.org/ftp/arxiv/papers/1703/1703.00924.pdf

Step 2: Identifying Key Applications

Application usage profiles have been measured for leading sites around the globe by hour and number of systems used.
For each application:

Application system utilization = (app time [hrs] x # nodes used) / (total # nodes x total time [hrs])

In most cases, a small number of applications account for a very large percentage of the overall demand. In fact, the top 15 applications typically drive over 75 percent of the system utilization. Recognizing this usage profile allows a site to focus on a much smaller set of applications to understand their current and forecasted usage trends.

Figure-2: Histogram of typical HPC system utilization by application

Understanding the workload profile identifies the key applications and helps define the eventual system configuration. A quick glance at the application profile for these three sites illustrates that 80 percent of the key applications are currently GPU-accelerated. The overall throughput for the entire system would benefit by accelerating all of the key applications.

Step 3: Exploring New Applications

The system you deploy today will have to live for three to five years. Therefore, you shouldn't just look at today's workloads. To get further insight about system needs, it's important to understand what key users and potential new users are planning to run in the future.

There's an emerging trend toward using data science in addition to traditional computational science to maximize throughput. Data science is a method of extracting knowledge and insights from data in various forms, whether measured or produced by simulations. This employs techniques and theories drawn from broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning. GPU accelerators have demonstrated great success with machine

learning and are the best platform for this new application trend, as the computational needs for data science applications will quickly overwhelm a CPU.

Traditionally, outfitting a data center for great science required building the biggest system possible comprised of a lot of computational nodes. But arbitrarily adding compute nodes won't equate to improved throughput for future workloads. New HPC data center systems require a design that's built to anticipate both the current and future application needs of its users. Profiling application usage helps identify how the HPC data center is used today, but it doesn't anticipate future system requirements. The best way to anticipate future requirements is working directly with scientists and developers to understand what they're doing next.

Traditional HPC applications continue to be modified to take advantage of parallel processing and threading to improve throughput and accuracy. In fact, over 550 applications (including all of the top 15 HPC application) are GPU-accelerated. A well-designed HPC system built to solve next-generation scientific challenges needs to include the right amount of GPU accelerators to allow researchers to innovate faster while reducing data center costs.

Deploying Accelerators to Maximize Throughput

First, it is important to realize that the benefit delivered by a supercomputer is measured in how much science can be performed over time and not based on a High-Performance Linpack (HPL) score. GPU computing has reached a tipping point in the HPC market that will encourage continued application optimization. Intersect360 Research recently published a **paper**[4] that listed the 50 most commonly used HPC applications. According to the latest HPC User Site Census data and additional research, over 70 percent of the 50 most popular application packages mentioned by HPC users offer GPU support. All of the top 15 applications currently have some form of GPU support.

Recently, three supercomputing sites—Tsubame 3, Piz Daint, and Summit—performed a workload analysis and determined the need to deploy GPU-accelerated nodes to meet the demands of their customers. Although each concluded that GPUs are key, they all have deployed GPUs a little differently (see table below). In all three cases, the target workload profile included traditional HPC applications and AI.

4. paper: http://www.nvidia.com/content/intersect-360-HPC-application-support.pdf

	PIZ DAINT (2016)	TSUBAME 3 (2017)	SUMMIT (2018)
Compute Node Configuration	Mix of CPU-only and GPU nodes 5,320 XC50 nodes with one P100 GPU: two Xeon 2690 CPUs 1,431 XC40 nodes with two Xeon 2695	All GPU-accelerated nodes 540 nodes with two Xeon 2680 v4 CPUs and four P100 GPUs	All GPU-accelerated nodes 4,600 nodes with Power9 CPUs: six V100 GPUs
Performance (Peak PF)	25	12	200
Power (MW)	2.3	12	13
Networking	Aries	Omnipath	InfiniBand
Target Workload	Primary workload: traditional HPC + emerging AI user	Mixed workload: traditional HPC + AI and ML	Heavy emphasis on large-scale science (including **13 CAAR applications**[5]) + AI
Power (MW)	2.3	12	13

Throughput Analysis

NVIDIA provides customizable workload analysis tools that can help customers optimize their data center throughput and costs savings. Using data center budget and application workload information as inputs, these tools can calculate cost savings and throughput improvements by deploying GPU-accelerated nodes. Whether the recommended configuration includes a large number of GPU nodes depends on if GPU-accelerated applications are a part of the application workload mix. The main objective of this throughput analysis is to make designing your future data center simple and effective.

Conclusion

HPC data centers need to be built to maximize throughput and enable more science achievements in a shorter amount of time. Maximizing throughput isn't only about building the biggest, fastest system but also about anticipating future application workload demands. To ensure the data center is designed to provide the maximum throughput, it's critically important to first understand the workload profile of the data center users. As was shown in the workload analysis completed by supercomputing sites, about 15 applications typically consume 75 percent of the system workload. Profiling past application workloads is a good first start, but future HPC data centers will need to be built for both traditional computational science and emerging data science. Adding the power of data science (AI) to the already powerful HPC data center will lead to science achievement once thought impossible. The HPC data center of the future will require the right amount for GPU

5. Source: https://www.olcf.ornl.gov/caar/

accelerators to maximize throughput, reduce costs, and increase the amount of science it can achieve.

REFERENCES:

(1) HIV capsid
Researchers at the University of Illinois at Urbana-Champaign (UIUC) leveraged the application acceleration made possible by HPC systems to achieve a major breakthrough in the battle against HIV by uncovering details about the structure of the HIV capsid. The work centered on accelerating NAMD (Nanoscale Molecular Dynamics) using the National Science Foundation's (NSF) petascale computing system Blue Waters. GPUs provided the scalable performance NAMD needed to determine and illustrate the structure of the HIV capsid. This enabled researchers to discover potential vulnerabilities that could be exploited, which led to the development of new drugs that effectively fought the HIV virus by targeting its capsid.

(2) Weather prediction
Predicting the weather is a high-stakes game. Ten years ago, Hurricane Katrina devastated New Orleans. Three years ago, Hurricane Sandy battered New York City. Hundreds lost their lives, and damages were in the billions. Researchers at the Swiss Federal Office of Meteorology and Climatology, **MeteoSwiss**[6], determined that they needed to create weather models for short-range (24-hour) forecasts from 2.2 km to 1.1 km resolution. A powerful new system could forecast the amount, duration, and location of rain and snow in more detail and provide early warning forecasts for severe weather events. MeteoSwiss deployed the GPU-accelerated version of the **COSMO**[7] model, which is extensively used by other national weather services in Germany, Italy, Greece, Poland, Romania, and Russia and in regional climate studies at more than 70 research institutes. They developed a new system that provides 40 times higher performance than the CPU-based system it's replacing. It allows MeteoSwiss to develop weather models with more than two times higher resolution—and three times higher energy efficiency. This makes MeteoSwiss the first major national weather service to deploy a GPU-accelerated supercomputer to improve its daily weather forecasts.

(3) Accelerating drug discoveries
To reduce the computation time and hasten the discovery of potentially life-saving anti-cancer drugs, Science Applications International partnered with Silicon Informatics to speed up the processes that are used to identify the small molecules that inhibit cancerous diseases,. The faster these molecules are identified, the more quickly the National Cancer Institute (NCI) can study them and create cancer-fighting drugs. Silicon Informatics software kernels, together with NVIDIA® Tesla® GPUs, allowed Science Applications International Corporation (SAIC) to speed up this process by a factor of 10. This means that a scientist who had to wait overnight to receive results can now work more interactively, performing several operations during a single work day.

6. MeteoSwiss: http://www.meteoswiss.admin.ch/home.html?tab=overview

7. COSMO: http://www.cosmo-model.org/content/default.htm