The Increasing Fusion of HPC and Large-scale Data Analytic Workloads

Bill Blake
Exec VP R&D

26th International Symposium on Computer Architecture and High Performance Computing
SBAC-PAID
October 22-24, 2014
Paris, France
Outline

• Key Trends in HPC and Big Data

• Technology Factors in “Convergence”

• Technology Factors in Storage and Memory

• Implications to Future System Architectures

• Application Focused Hardware/Software Stacks (“Appliances”)

2
The Fusion of Supercomputing with Large Scale Data Analytics

Key Trends

The cost per FLOP is dropping to near zero!

Application needs for memory size and IO performance are insatiable!

New levels of memory and storage hierarchy are emerging due to Non Volatile Memories

MPP interconnects must go beyond MPI and support atomic Global Memory Operations

Future systems are being influenced by the Converged Infrastructure of compute & storage

Converged infrastructure demands management of O(10000) nodes for Exascale or Hyperscale
Exascale and Hyperscale
Two Paths to Millions of Nodes and Billions of Threads

Supercomputing (Exascale)
- Top-down model driven
- Scalable computing w/high BW, low-latency, Global Memory Architectures
- Highly integrated CPU/GPU-memory-interconnect & network storage
- Highest Bandwidth, Lowest Latency access to single name space
- Minimize data movement – load the “mesh” into memory
- Move data only for loading, check-pointing or archiving
- “Basketball court sized” systems

Cloud Computing (Hyperscale)
- Bottom-up data driven
- Distributed computing at largest scale
- Divide-and-conquer approaches on Service Oriented Architectures
- Latency tolerant algorithms with IP network access
- Maximize data movement-- Scan/Sort/Stream all the data all the time
- Lowest cost processor-memory-interconnect & local storage
- “Warehouse sized” systems
The Fusion of Supercomputing and Big & Fast Data

Modeling The World
HPC solving “grand challenges” in science, engineering and analytics

Data Models
Integration of datasets and math models for search, analysis, predictive modeling and knowledge discovery

Data-Intensive Processing
High throughput event processing & data capture from sensors, data feeds and instruments

Math Models
Modeling and simulation augmented with data to provide the highest fidelity virtual reality results

Bottom Up Data Driven Latency tolerant
Performance metric: time to insight and time to ingest data
Future: more predictive workloads

Top Down Model Driven Latency intolerant
Performance metric: FLOPS and time to output data
Future: more data assimilation
Need to Create a “Virtuous Cycle”

Cloud provides new distributed programming models that utilize “divide and conquer” approaches with massive scale-out Service Oriented Architectures using local storage and low cost hardware, and new data analytics algorithms where data scientists claim “the larger the data the simpler the algorithm”

HPC provides new parallel programming models that utilize highly scalable Global Memory Architectures supported by highest BW, lowest latency interconnects, with powerful algorithms for high fidelity modeling and simulation using highly iterative processing of both capability and capacity workloads that increasingly support data assimilation (from sensors)
The “Fusion” Challenge

Supercomputing minimizes data movement – “data movement” is highly restricted for defensive or resiliency such as loading, check pointing or archiving. Programming model is imperative (C++/Fortran + MPI) with focus on the details of how parallel programming is done.

Data-intensive computing is all about data movement - scanning, sorting, streaming and aggregating all the data all the time to get the answer or discover new knowledge from unstructured or structured data sources. Programming model is declarative (query) or functional with emphasis on what is being computed versus how it is computed.

Cloud Computing is all about virtualization -- Application access to converged infrastructure (Compute/Network/Storage) where the focus is on what is being computing rather then where the computing is done.
# Breathtaking Changes Resulting in “Free FLOPS”

## TALE OF THE TAPE: SUPERCOMPUTER VS. GAME CONSOLE

<table>
<thead>
<tr>
<th></th>
<th>SANDIA LAB’S ASCI RED</th>
<th>SONY PLAYSTATION 3</th>
<th>SANDIA LAB’S RED STORM XT</th>
<th>NVIDIA-powered Game Console</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DATE OF ORIGIN</strong></td>
<td>1997</td>
<td>2006</td>
<td>2005</td>
<td>2019</td>
</tr>
<tr>
<td><strong>PEAK PERFORMANCE</strong></td>
<td>1.8 teraflops</td>
<td>1.8 teraflops*</td>
<td>40 teraflops</td>
<td>40 teraflops²</td>
</tr>
<tr>
<td><strong>PHYSICAL SIZE</strong></td>
<td>150 square meters</td>
<td>0.08 square meter</td>
<td>280 square meters</td>
<td>0.08 square meter</td>
</tr>
<tr>
<td><strong>POWER CONSUMPTION</strong></td>
<td>800 000 watts</td>
<td>&lt;200 watts</td>
<td>1 000 000 watts</td>
<td>&lt;200 watts</td>
</tr>
</tbody>
</table>

* For GPU; CPU adds another 0.2 teraflops

² for GPU; CPU adds another ~10 teraflops (10nm process)
Big Data Means New Kinds of Data

Big Data refers to data that is not easily captured, managed and analyzed by traditional tools due to its higher volume, velocity, and most importantly variety.

Science will increasingly be data-driven to understand the world
Business will increasingly be data-driven to understand customers

Source: IDC White Paper sponsored by EMC May 2009
Big Data → Fast Data

SUPERCOMPUTERS & ANALYTIC APPLIANCES

Virtualized Converged Infrastructure

CLOUDS

Distributed Memory

GRID
LAN/WAN interconnects

SAN Interconnects

Enterprise Data (structured)

Ethernet Clusters

InfinBand Cluster

MPP Global Memory

SMP Shared Memory

Fast Data

Information Density & MPP Integration and Memory Increase

Data Volume & Interconnect latency Increase
Networking Bandwidth is Outpacing Server Bandwidth

A “virtuous cycle” happens as higher bandwidth encourages convergence of compute/network/storage with virtualization, big data, and cloud adoption moving workloads away from dedicated servers which in turn demands more network bandwidth …

- Network Bandwidth doubling every 18 months or less
- Server Bandwidth doubling every 24 months
- Intel, IBM and HP are developing silicon photonics where chips will emit light that will support 100Gbps optical interconnects both within the rack and the datacenter at large – A potential game changer!
- Quanta has demonstrated a rack of optically connected CPUs in a tray where an upgrade with new CPUs can be swapped without the need for a entire new motherboard or system

Silicon Photonics will accelerate this trend creating the need for innovation in system fabric and datacenter infrastructure
The Future: Global Memory and Latency Hiding


Data Reuse Near Previous Data Access

Trend

Modeling & Simulation (dense matrices)

Informatics – Analytics (sparse matrices)

Data Reuse over Time
Big Data ➔ Fast Data Programming Emphasis

“what” plus “how”
Highly optimized C++/Fortran Thread/Task + Msg Parallelism Increasingly Auto-tuned Tools expose low level

“what” without “how”
Functional and Declarative Languages

“what” without “how” or “where”
PaaS

SUPERCOMPUTERS & ANALYTIC APPLIANCES
Virtualized Converged Infrastructure
CLOUDS

“what” plus “how”
Highly optimized C++/Fortran Thread/Task + Msg Parallelism Increasingly Auto-tuned Tools expose low level

“what” without “how”
Functional and Declarative Languages

“what” without “how” or “where”
PaaS

SUPERCOMPUTERS & ANALYTIC APPLIANCES
Virtualized Converged Infrastructure
CLOUDS
The Next Generation of Analytics Will Be Enabled By Very High Performance Global Memory Operations

- **SMP Shared-memory**
- **MPP Global Memory**
- **ISupercomputing Cluster**
- **SAN Interconnects**
- **Enterprise Data (structured)**
- **GRID LAN/WAN interconnects**
- **Big Data**
- **Etternet Clusters**

**Capability Workloads**
- + 5 Years
  - Converged MPP Petascale Global Memory Architecture for Analytics

**Capacity Workloads**
- Ingest and Archive
- High Throughput, Multi-tenancy
- Cloud Computing
- The last bastion of Magnetic Storage?
The Future of Magnetic Storage

- Processors have increased in speed 200,000X since 1970

- But spinning mechanical disk drives have decreased latency by only 9X (takes ~8 msec for a full rotation after a miss)

- Streaming bandwidth has increased 193X but still involves serial bits passing under the R/W heads

- Disk firmware has grown to the size of an O/S due to defensive steps needed to enable a four year warranty

- But areal density improvements, a traditional metric for HDD capacity and ultimately price, continue to double every year 1,000,000X since 1970 (35 million times since 1956!)
Seagate Turns On the Heat!

- Future HDDs will use Heat Assisted Magnetic Recording (HAMR) to make it easier to write data at higher densities, reaching 10TB/in\(^2\)
- Bit Patterned Media using chip lithography techniques for smaller spots will combine with HAMR further increasing density to 50TB/in\(^2\)
- 50 TB/in\(^2\) density would provide enough storage to put the Library of Congress on a single HDD
- Beyond 100 TB/platter by 2020, the next generation of NVM SSDs wins
DRAM reaching scaling limits

- DRAM technology faces major scaling challenges (e.g. leakage) as feature sizes shrink further to 10 nm
- DRAM power dissipation is growing and will reach 40% of total system power
- New non-volatile memories are reaching the scale of DRAM (without the dynamic and static power problems)
- The downside? Cost and limited life
Storage Class Memories

- Solid-state memory in the boundary between storage and memory – post NAND and DRAM

- A definition per IBM Almaden researchers:
  - No more than 3X the cost of enterprise HDD
  - <200 nsec Read/Write/Erase time
  - >100,000 Read I/O operations per second
  - >1 GB/sec Bandwidth
  - Lifetime of $10^8$ to $10^{12}$ write/erase cycles
  - 10 X lower power than enterprise HDD
System Implications of Fast, Cheap, Non-Volatile Memory

• Operating System
  – **Evolutionary changes** integrated into existing architectures (same file system semantics but with buffering, persistent objects)
  – **Revolutionary changes** with whole-system persistence will effect memory management, I/O, fault management, etc.
  – Relegate magnetic storage to an archival function

• Power
  – Today 30% to 40% of system power is DRAM (HDD is 10%)

• Applications
  – Manipulate data 100x – 1000x faster (both throughput & latency improve)
NVM-based Architectural Possibilities

- **Compute Intensive**
  - CPUs
  - DRAM
  - IO
  - HDD

- **Data Intensive**
  - CPUs
  - DRAM
  - SSD

- **“Storage Hierarchy”**
  - CPUs
  - DRAM
  - IO
  - HDD

- **“No Moving Parts”**
  - CPUs
  - DRAM
  - SSD

- **“Active NVM”**
  - CPUs
  - DRAM
  - CPU + IO
  - HDD

- **“Stacked CPU/NVM Chip?”**
  - CPUs
  - NVM
  - SOC
  - CPU
  - SSD
  - Net
Evolution of Big Data

Initially, Transaction Analytics (OLAP) using ad hoc SQL queries on structured data in relational databases by analysts produced BI Reports.

Looking at all the data, O(100) TB, all the time.

Recently, Textual Analytics (MapReduce) provided an API for analysis of unstructured data in massive data sets by programmers seeking “long tail” insights.

Looking at all the data, O(1000) TB, once at a time.

And based on new standards, Graph Analytics (RDF, OWL) with ad hoc SPARQL queries on linked data by analysts seeking discovery via hypothesis.

Looking at all the data, O(100) TB, and relationships.
### Analyzing Analytics: A “Swim Lane” View

<table>
<thead>
<tr>
<th>Key Function</th>
<th>Language</th>
<th>Data Approach</th>
<th>SMP Server</th>
<th>Cluster And MPP</th>
<th>Cloud And Grid</th>
<th>Web Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Oracle</td>
<td>OLTP</td>
<td>Declarative (SQL)</td>
<td>Structured (relational)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IBM (NZ) Teradata</td>
<td>OLAP Ad Hoc</td>
<td>Declarative (SQL+UDF)</td>
<td>Structured (relational)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cray?</td>
<td>Semantic Ad hoc</td>
<td>Declarative (SPARQL)</td>
<td>Linked, Open (graph-based)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon Google</td>
<td>OLAP Ad Hoc</td>
<td>Procedural (MapReduce)</td>
<td>Unstructured (Hadoop files)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td>Optimize Models</td>
<td>Procedural (Solver Libs)</td>
<td>Optimization &lt;-&gt; Simulation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cray IBM</td>
<td>Simulate Models</td>
<td>Procedural (Fortran, C++)</td>
<td>Matrix Math (Systems of Eq’s)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Decision Making through Fact Finding and/or Equation Solving

<table>
<thead>
<tr>
<th>Key Function</th>
<th>Language</th>
<th>Data Approach</th>
<th>“Airline” Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLTP</td>
<td>Declarative (SQL)</td>
<td>Structured (relational)</td>
<td>ATM transactions Buying a seat on an airplane</td>
</tr>
<tr>
<td>OLAP Ad Hoc</td>
<td>Declarative (SQL+UDF) or NoSQL</td>
<td>Structured (relational)</td>
<td>Business Intelligence analysis of bookings for new ad placements or discounting policy</td>
</tr>
<tr>
<td>Semantic Ad hoc</td>
<td>Declarative (SPARQL)</td>
<td>Linked, Open (graph-based)</td>
<td>Analyze social graphs and infer who might travel where</td>
</tr>
<tr>
<td>OLAP API for Ad hoc</td>
<td>Procedural (MapReduce)</td>
<td>Unstructured (Hadoop files)</td>
<td>Application Framework for large scale weblog analysis</td>
</tr>
<tr>
<td>Data Assimilation</td>
<td>Procedural (C++, Fortran)</td>
<td>Data merged With simulations</td>
<td>Sensor data incorporated into computer simulation</td>
</tr>
<tr>
<td>Optimize Models</td>
<td>Procedural (Solver Libs)</td>
<td>Optimization &lt;-&gt; Simulation</td>
<td>Complex Scheduling Estimating empty seats</td>
</tr>
<tr>
<td>Simulate Models</td>
<td>Procedural (Fortran, C++)</td>
<td>Matrix Math (Systems of Eq’s)</td>
<td>Mathematical Modeling and simulation (design airplane)</td>
</tr>
</tbody>
</table>
Declarative Versus Imperative (Language)

**SQL**

```
SELECT list FROM table
WHERE condition
GROUP BY category
```

**MapReduce**

Map() function selects data
<There are no tables>
Reduce() handles grouping
Is There A Better Way to Express Analytics? Queries and Program Code Do Not Mix Well

Language close to the machine

Problem Domain Abstraction

Language close to the human

Domain Expert

Expert Programmer

Analyst

Excel

Mathematica (Symbolic)

MATLAB

Julia

NumPy

Matrix/Array Languages

Fortran

C++

C

C#

Python

Java

Scala

Erlang

F#

Chapel

UPC

KDT

NoSQL

SQL

SPARQL

Query Languages

Domain

Expert

Programmer

Analyst

Is There a Better Way to Express Analytics?

Queries and Program Code Do Not Mix Well
Software-led Enterprise Infrastructure

There are substantial changes in the technology used by Datacenters that Supercomputing users must understand and adopt to support data analytics (very different from Exascale!)

- Need to focus on applications and workloads and move to converged infrastructure
- Appliances are first stage of converged infrastructure (compute + storage)
- Cloud is the ultimate converged infrastructure with virtualized integration of fabric, blades & storage
- Software-defined Storage and Network then Software-defined Data Center

Source: Booz Allen Hamilton “Concepts in the Cloud”
Virtual Machine Images and the Cloud

• Originally, virtualization was introduced to run more applications on fewer servers (addressing the <10% utilization issue of servers with apps waiting)
• With fast networks a new application and O/S VMI can be loaded in seconds
• The next step was moving VMIs to the Cloud
  – Amazon was quick to offer entire MPI clusters as VM Images to run on the Cloud
  – Amazon announced recently a parallel RDBMS with big memory and solid state drives as a VMI!

Largest Amazon run (so far)¹:

156,000+ cores measured at 1.21 PetaFLOPS peak throughput (not RMax) Top10 starts at 2.9 PFLOP

264 years of computing in 18 hours running all eight AWS regions (over five continents)

Rather than spending $68 Million to buy this system, workload completed for $33,000

¹ Organic semiconductor simulation - http://blog.cyclecomputing.com/amazon-ec2/
Power Consumption limits Exascale and Hyperscale

• Redundant Arrays of Inexpensive Nodes sounds cool but it is anything but cool!
• Google runs over 1,000,000 servers (> 260 MWatt!)
• 100 searches = power to iron 1 shirt!
• The fusion of HPC and Large Scale Data Analytics will only increase power demands

Ref: Rich Miller blog 2011
Limits of General Purpose Systems

Parallel and distributed computing techniques now deliver price/performance gains in excess of Moore’s Law in standard systems, however...

Gains from application specific architectures can exceed 10X to 100X when the system is designed to fit the application workload.
The Appliance vs. General Purpose System

- Appliance concept started with Cisco and NetApp
- Netezza Data Warehouse Appliance (FPGA-based)
- Maxeler Technologies Data Flow Appliance (FPGA-based)
- Cray Urika Graph Appliance (ASIC-based)
- DE Shaw Anton Molecular Dynamics Appliance (ASIC-based)
- D-Wave Adiabatic Quantum Annealing Appliance (SQUID-based)
- Dramatic 1000X+ price/performance gain is the key motivator
- Architectural continuum? General Purpose -> Appliance -> SPD
The Netezza (now Puredata) Approach to Appliances: Moving Processing to the Data

Active Disk architectures

- Integrated processing power and memory into disk units
- Scaled processing power as the dataset grew

Decision support algorithms offloaded to Active Disks to support key decision support tasks

- Active Disk architectures use stream-based model ideal for the software architecture of relational databases

Influenced by the success of Cisco and NetApp appliances, the approach combined software, processing, networking and storage leading to the first database warehouse appliance!

Netezza is an IBM Company

Streaming Data Flow Unclogs the Bottleneck

IBM/Netezza Performance Server

Linux SMP Host

Postgresql RDBMS is sole application running on the server with complete control over active disk storage nodes → DW appliance

Key to performance of ad hoc queries on the VLDB was delivering the highest bandwidth access to tables on disk
Merging Processing and Storage Devices
The Interesting Case of Nimble Storage

A Flash/HDD storage vendor that built a media-agnostic new file system and an application centric way to sell devices

Nimble Storage delivers 50X advantage!

A great example of a software savvy approach to application needs based on a blend of hardware components that combine the best features of Flash & HDD
New Active Storage Ideas

Very Interesting research has been published from HP Labs, ORNL, U Michigan and Georgia Institute of Technology on future Data-Centric system architectures:

– Utilization of “active” NVRAM in storage hierarchy
– On-line processing and “data staging” where I/O and data movement actions are enhanced with computation to better filter, reduce, sort, compress data.
– Applications running with this “offload” have shown 3X to 4X speedups
– The next thing after “active” Disks?

Proceedings of the 2nd international workshop on Petascale data analytics: challenges and opportunities
Dealing with High Dimensionality Data “déjà vu all over again!”

Ref: Dr. Paul Messina, “Enabling Technologies for Beyond Exascale Computing”
High Performance Computing Workshop, Cetraro (Italy) July 2014
A new “Appliance” from D-Wave Systems

• World’s first commercial quantum computing company

• Mission: To help solve the most challenging problems in the universe
  – Focus: Optimization and Machine Learning

• 512 qubit systems installed at
  – Lockheed Martin /USC
  – Google/NASA Ames

• 100 U.S. patents, 70+ papers published
What is a Quantum Computer?

• Exploits quantum mechanical effects
• Built around quantum bits “qubits” rather than “bits”
• Operates in an extreme environment
• Enables quantum algorithms to solve very hard combinatoric problems that are very difficult to parallelize in machine learning and optimization
• Directly analyze data with very high dimensionality
Processor Environment

- Cooled to 0.02 Kelvin, 150x colder than interstellar space
- Shielded to 50,000x less than Earth’s magnetic field
- In a high vacuum: pressure is 10 billion times lower than atmospheric pressure
- On low vibration floor
- 192 i/o and control lines from room temperature to the chip
Programming Environment

• Operates in a hybrid mode acting as a co-processor or accelerator
• D-Wave system is “front-ended” by a standard server on a network
• D-Wave system executes one quantum machine instruction
  – User formulates problem specifying the weights at each data point and interaction strengths between the points
    – Specify up to ~500 qubit weights
    – ~1500 interaction strengths
    – # of solutions
• Performs ~10,000 solutions per second
• Results are returned to the front-end and back to the user
Quantum Effects

Superposition

Entanglement

Quantum Tunneling
Quantum Annealing

- Classical computers running classical algorithms can only walk over this landscape.
- Quantum computers can tunnel through the landscape making it faster to find the lowest point using quantum annealing.
- Quantum annealing “tunes” qubits from superposition state to classical state to return a set of answers scored to show the best solution.
Processing Using D-Wave

- Made from a lattice of tiny superconducting circuits (qubits)
- Chilled close to absolute zero to get quantum effects
- User maps a problem into search for “lowest point in a vast landscape” which corresponds to the best possible outcome
- Processor considers all possibilities simultaneously to determine the lowest energy required to form those relationships
- That state, seen as the optimal outcome, is the answer
No one jumps a 20 foot chasm in two 10 foot jumps.

Miguel Guhlin
Concluding Comments

- Cloud computing provides an excellent multi-tenancy resource for high throughput capacity computing especially where virtualization of converged compute/network/storage affords the use of resources in various locations.

- But highly parallel analytic workloads, especially those that require low latency messaging and/or global memory operations that benefit greatly from the high performance interconnects and tight integration of MPP machines, will not migrate from MPP to Cloud.

- Many Cloud developments will “condense” into future big memory MPP systems, including programming models, RDF and NoSQL databases, software defined networks and storage, and hypervisors that combined with the high performance message passing and global atomic memory support in MPP networks (e.g., Cray Aries and future Intel® Fabrics) will best support the fusion of HPC and large-scale analytics.

- New ‘Killer Apps” will emerge from research into computational intelligence (e.g., deep learning with high dimensionality data) that run best on optimized hardware/software stacks (appliances) or new approaches altogether such as quantum computing.
Thanks!

bill.blake@dwavesys.com