IBM Data Centric Systems & OpenPOWER

Yoonho Park
Research Staff Member/Senior Manager
Data Centric Systems Software, Cloud and Cognitive

IBM Research
Data Growth Outpaces Computing Technology Elements

Data volume grows exponentially

Microprocessor clock rates have stalled...

I/O performance/capacity loosing ground...

And network bandwidth cannot keep up.
Big Data and the New Era of Computing

Data volume is on the rise

- Sensors & Devices
- Social Media
- VoIP
- Enterprise Data

Dimensions of data growth

- **Volume**: Terabytes to exabytes of existing data to process
- **Velocity**: Streaming data, milliseconds to seconds to respond
- **Variety**: Structured, unstructured, text, multimedia
- **Veracity**: Uncertainty from inconsistency, ambiguities, etc.

- **Big Data analytics and Exascale High Performance Computing facing similar challenges**: scale, performance, bandwidth, computational complexity
- **IBM approach**: Move compute to data – Data Centric Systems (DCS)
Different Solutions for Different Types of Workflows

- Conceptual View of Data Intensive with Floating Point Workflow
- Conceptual View of Data Intensive-Integer Workflow
- Low Spatial Locality
- Integer OPS
- Conceptual View of Data Intensive with Floating Point Workflow

- Data Centric Applications
- Compute Centric Applications

- Floating Point OPS
- High Spatial Locality

Region defined by LINPACK
Data Centric Workflows: Mixed Compute Capabilities Required

Analytics Capability:
- Complex code
- Data Dependent Code Paths / Computation
- Lots of indirection / pointer chasing
- Often Memory System Latency Dependent
- C++ templated codes
- Limited opportunity for vectorization
- Limited scalability
- Limited threading opportunity

Massively Parallel Compute Capability:
- Simple kernels,
- Ops dominated (e.g. DGEMM, Linpack)
- Simple data access patterns.
- Can be preplanned for high performance.
Comparing Compute Centric to Data Centric

- Systems and Solutions must become more data centric and data aware
  - Data movement minimized
    - Within the system
    - Within/across the end to end solution
  - Compute enabled at all levels
  - Workloads/workflow driven system and solution design choices
  - Modular, composable solution architectures
  - Enhanced resource agility and sharing

- Focus must shift from algorithms to workflows
  - New end to end efficiencies and optimizations
  - Based on a data aware understanding of the full scope of resource requirements
    - Storage, Networking, Compute, Applications, Resource management, Data Centers

- Uncertainty of the future puts a premium on flexibility and innovation
IBM Data Centric Design Principles

Massive data requirements drive a composable architecture for big data, complex analytics, modeling and simulation. The DCS architecture will appeal to segments experiencing an explosion of data and the associated computational demands.

**Principle 1: Minimize data motion**
- Data motion is expensive
- Allow workloads to run where they run best

**Principle 2: Enable compute in all levels of the systems hierarchy**
- HW & SW innovations to support / enable compute in data

**Principle 3: Modularity**
- Balanced, composable architecture for Big Data analytics, modeling and simulation

**Principle 4: Application-driven design**
- Use real workloads/workflows to drive design points

**Principle 5: Leverage OpenPOWER to accelerate innovation and broaden diversity for clients**
# IBM OpenPOWER-based HPC Roadmap

<table>
<thead>
<tr>
<th>Mellanox Interconnect Technology</th>
<th>IBM CPUs</th>
<th>NVIDIA GPUs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>POWER8</strong></td>
<td>Connect-IB FDR Infiniband PCIe Gen3</td>
<td>Kepler PCIe Gen3</td>
</tr>
<tr>
<td>OpenPower CAPI Interface</td>
<td><strong>POWER8+</strong></td>
<td>Pascal NVLink</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ConnectX-4</strong> EDR Infiniband CAPI over PCIe Gen3</td>
<td></td>
<td>Volta Enhanced NVLink</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>ConnectX-5</strong> Next-Gen Infiniband Enhanced CAPI over PCIe Gen4</td>
<td></td>
</tr>
</tbody>
</table>

**IBM Nodes**
- 2015
- 2016
- 2017

Heterogeneity is Key
OpenPOWER, a catalyst for Open Innovation

**Market Shifts**

- Moore’s law no longer satisfies performance gain
- Growing workload demands
- Numerous IT consumption models
- Mature Open software ecosystem

**New Open Innovation**

- Rich software ecosystem
- Spectrum of power servers
- Multiple hardware options
- Derivative POWER chips

---

**Open Development**

collaboration of thought leaders

**Open Software**

open software, open hardware

**Performance**

amplified capability

---

*OpenPOWER is an open development community, using the POWER Architecture to serve the evolving needs of customers.*
US & UK Research Centers Select OpenPOWER-based Supercomputers

IBM, Mellanox, and NVIDIA awarded $325M U.S. Department of Energy’s CORAL Supercomputers

CORAL: Leadership Class Supercomputers

5x – 10x Higher App Perf Than Current Systems

IBM & UK’s STFC in £313M Partnership for Big Data & Cognitive Computing Research
Hybrid CPU/GPU architecture

- At least 5X Titan / Sequoia Application Performance
- Approximately 3,400 nodes, each with:
  - Multiple IBM POWER9™ CPUs and multiple NVIDIA Tesla® GPUs using the NVIDIA Volta architecture
  - CPUs and GPUs completely connected with high speed NVLink
  - Large coherent memory: over 512 GB (HBM + DDR4)
  - All memory directly addressable from the CPUs and GPUs
- Over 40 TF peak performance per node
- Dual-rail Mellanox® EDR-IB full, non-blocking fat-tree interconnect
- IBM Elastic Storage (GPFS™) - 1TB/s I/O and 120 PB disk capacity.
Programming Approaches

- **Accelerator Approach:**
  - Required when not coherent
  - Each processor computes in its own private address space
  - Data objects are homed in CPU Memory, are copied to GPU memory for GPU execution, GPU engines act only on data in GPU memory.

- **Compute in Shared Address Space:**
  - New option, now that CPU / GPU memories are coherent
  - Data objects can be in any physical memory domain
  - Processors (either CPU or GPU) can use data in place.
  - No copies required

- **Note:**
  - Will still have to manage NUMA
Compute and Memory View – Emerging Approach

Multiple Compute Engines
• Consider all engines as equal peers

Multiple Memories
• Consider all memories as equal peers
Compute and Memory View – Traditional Acceleration

Data / Compute Flow

MPI Task

Data
Peer Processing

Data can be placed at Allocation, or can migrate under run time control (e.g. UVM)
Thread-like programming model (OpenMP 4.x, OpenAcc, CUDA …
Still under development …
Future DCS Programming Model

Big Data workflows
- Unstructured or semi-structured data
- Collaborative Filtering, clustering, web search, recommendation, …
- Real-time analysis, fraud and anomaly detections
- Sensor data filtering, classification, …
- Statistical averages/histogram, …
- Deep Learning, Support Vector Machine, …
- Graph Community finding to Shortest Path, …

HPC workflows
- Structured data
- Exascale data sets
- Primarily scientific calculations
- Ensemble analysis
- Sensitive analysis
- Uncertainty quantification

Common challenges
- Network bandwidth/efficiency
- CPU clock rate stalling
- IO performance
- Layered storage: New storage technologies

HPC ecosystem
- Fortran/C, MPI+OpenMP, CUDA, Legion, …
- UPC, OpenSHMEM, Charm++, GASNet, …
- UCX, PAMI, Verbs, …

Exascale Runtime & Middleware
- High productivity, fault-tolerance

1GigE, TCP/IP, HDDs, …

Commodity clusters
- High productivity and High performance

OpenPOWER/DCS
- High performance, specialized hardware

CAPI, InfiniBand, Flash

© 2016 IBM
DCS OpenPOWER Software Contribution Examples

- Linux – NUMA support for hardware coherent GPU memory
- Provisioning – xCAT
- Burst Buffer – Support for fast shared-file checkpointing
- CSM – Cluster System Management
- Compilers – LLVM
- Tools – Ensure tools have appropriate APIs
Experience/Observation: Extraction of meaningful insights from Big Data and enabling real-time, predictive decision making requires similar computation techniques that have been characteristic of Technical Computing

- Convergence in many future workflow requirements including Big Data-driven analytics, modeling, visualization, and simulation
- Will require optimized full-system design and integration

IBM Approach: Data Centric innovation in multiple areas, in open ecosystems with workload-driven co-design

- System architecture and design with modular building blocks
- Hardware technologies
- Integration of heterogeneous compute elements
- Software enablement