HPC and HPDA towards Cognitive, AI and Deep Learning

Traditional High Performance Computing
- Large Scale High Performance (CORAL) Solution
- IBM Power HPC Stack Cluster Solution

High Performance Data Analytics
- Big Data Strategy on Power
- HDP + Spectrum Scale
- HDP + Power
- Conductor with Spark

Deep Learning
- AI / Deep Learning Strategy for Power
- Power AI Platform
Our Compute System direction is based on Heterogeneous Processing and Big Data

Heterogeneous core types allow for optimizations in
• Power and chip area used
• Computation methods to address serial and parallel regions
• Power-efficient, frequency-optimized cores can be used for parallel regions while high-performance cores can be used for serial sections

Different memory characteristics can be optimized for the computational approach
• High BW or Low Latency or Energy Efficient
• Design focus on common shared memory for programmability and data access

Big Data drives the need for large memory and large BWs to Data for all compute units
IBM Power Systems: Open to the Core

<table>
<thead>
<tr>
<th>OpenPOWER</th>
<th>Open Source Workloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;300 members</td>
<td>mongoDB, Spark, hadoop, POSTGRES, cassandra</td>
</tr>
<tr>
<td>redhat</td>
<td></td>
</tr>
<tr>
<td>Mellanox</td>
<td></td>
</tr>
<tr>
<td>XILINX</td>
<td></td>
</tr>
<tr>
<td>NVIDIA</td>
<td></td>
</tr>
<tr>
<td>AMD, HP</td>
<td></td>
</tr>
<tr>
<td>Google</td>
<td></td>
</tr>
<tr>
<td>Dell</td>
<td></td>
</tr>
</tbody>
</table>

Expanding the ecosystem

OpenCAPI

Open Frameworks

torch, theano, Chainer, TensorFlow, Caffe
Leveraging OpenPOWER for CORAL Hardware

**IBM Power Systems S822LC (Minsky)**

- Up to 5.2 Tflops/GPU
  - Stacked Memory for increased BW, capacity & energy efficiency
  - Enhanced Unified Memory

- NVLink-1 5x faster than PCIe Gen 3
- Up to 12 SMT8 cores
  - CAPI Acceleration
  - Adaptive Power Management

**OpenPOWER**

- Up to 7.8 TF/GPU
  - Next Generation High Bandwidth Memory
  - Memory coherency

- NVLink-2 7-10x faster than PCIe Gen 3
- NVLink-2 7-10x faster than PCIe Gen 3
- Up to 24 SMT4 cores
  - CAPI v2, PCIe Gen 4
  - Superior Core Performance

**Mellanox Technologies**

- 100 Gb/s EDR
  - In-Network Computing with SHARP
  - Adaptive Routing

- Billion Cell Reservoir Simulation in record time (92 mins vs 20 hours)
- ResNet-50 90-epoch training in lowest time (7 hours vs 10 days) with highest accuracy (33.8% vs 29.8%)

**IBM Power Systems**

- POWER9 server for HPC & AI
- IBM Power Systems AC922 (Newell)
- Native RDMA
  - NVMe over Fabrics offload
  - PCIe Gen 4
  - CAPI v2 for fast virtual RDMA support
  - Hardware Tag Matching (automate pt-2-pt communication)
  - MPI rendezvous protocol offload
  - Precision time protocol support

**GPU can access CPU’s page tables**
Software Strategy Builds on CORAL Collaboration, Open Source, and OpenPOWER

Provide a complete HPC Software Solution for Power Systems
- Fully Integrated, Tested and Supported
- Exploit Power hardware features
- Focus on maintainability and reliability

- Open Source
- Supports multiple HW and Installation options

- Asynchronous Checkpoint transfer from node NVMe to file system
- Stage in data to expedite job start
- Stage out data releases CPU to start new job
- Flash wear, health and usage monitoring
- Open Source in 2018

- Single Namespace up to 250 Petabytes
- 2.5 TB/s large block sequential IO performance
- 2.6M file creates/sec for 32KB files in unique directories
- 50K file creates/sec to single shared directory

- Tuned for each Power Architecture
- Serial, SMP (OpenMP), GPU (CUDA), SIMD (VSX) and SPMD (MPI) subroutines
- Callable from Fortran, C, and C++ Parallel ESSL SMP Libraries

- New System Management SW Developed for CORAL
- Integrates system management functions
- Integrates job management
- Jitter aware design
- Data collection aligned to system clock
- Big Data for RAS analysis
- Open Source in 2018

- Hardware discovery & provisioning
- New System Productivity Feature for CORAL
- Job and Workflow Management
- Job and Workflow Management
- Job and Workflow Management

- Spectrum Scale
- Integrated File System
- Spectrum Scale
- Integrated File System
- Spectrum Scale
- Integrated File System

- ESSL/PESSL
- Engineering & Scientific Subroutine Libraries
- ESSL/PESSL
- Engineering & Scientific Subroutine Libraries
- ESSL/PESSL
- Engineering & Scientific Subroutine Libraries

- Compilers
- CUDA OpenMP 4.x
- OpenACC
- Compilers
- CUDA OpenMP 4.x
- OpenACC
- Compilers
- CUDA OpenMP 4.x
- OpenACC

- Optimize MPI Solution
- IBM Power Parallel Performance Toolkit for analyzing application performance including single timeline tracking for CPU, GPU & MPI

- Based on Open MPI
- New Features driven by CORAL Innovation including enablement of unique hardware features
- High Performance Collective Library
- Enhanced multithreaded performance
- IBM Power Parallel Performance Toolkit for analyzing application performance including single timeline tracking for CPU, GPU & MPI

- Rich collection of compiler supporting multiple programing models
- Optimized for Power
- Open Source and fully supported proprietary compiler options

- xCAT
- Hardware discovery & provisioning
- xCAT
- Hardware discovery & provisioning
- xCAT
- Hardware discovery & provisioning

- CSM
- New System Management SW
- CSM
- New System Management SW
- CSM
- New System Management SW

- Burst Buffer
- New System Productivity Feature for CORAL
- Burst Buffer
- New System Productivity Feature for CORAL
- Burst Buffer
- New System Productivity Feature for CORAL

- Spectrum LSF
- Job and Workflow Management
- Spectrum LSF
- Job and Workflow Management
- Spectrum LSF
- Job and Workflow Management

- PGI
- Rich collection of compiler supporting multiple programing models
- PGI
- Rich collection of compiler supporting multiple programing models
- PGI
- Rich collection of compiler supporting multiple programing models

- CUDA
- OpenMP 4.x
- OpenACC
- CUDA
- OpenMP 4.x
- OpenACC
- CUDA
- OpenMP 4.x
- OpenACC

- Enhanced job feedback for users, administrators
- RESTful API’s to simplify integration into business processes
- Absolute priority scheduling
- Enhancements for advance reservations
- Pending job limits

- Asynchronous Checkpoint transfer from on node NVMe to file system
- Stage in data to expedite job start
- Stage out data releases CPU to start new job
- Flash wear, health and usage monitoring
- Open Source in 2018
CORAL HPC System Overview

**Compute Rack**
- Over 850 Tflops peak performance per rack
- Multiple programing models & supported development environment
- Open Source Friendly
- Power Efficient
- Minimizes Jitter Impacts
- Air and Water cooled versions
- Up to 6 GPUs/server
- Up to 44 POWER 9 cores

**Login/Launch Servers**
- IBM Spectrum Scale RAID with declustered erasure coding
- Single Namespace up to 250 Petabytes
- 2.5 TB/s large block sequential IO performance
- 2.6M file creates/sec for 32KB files in unique directories
- 50K file creates/sec to single shared directory
- Spectrum Scale RAID with declustered erasure coding

**High Speed InfiniBand Network**

**Gateway Server**

**Workload Manager Server**

**Management Server**

**Service Server**

**Big Data System**

**IBM Power Systems POWER9 server for HPC & AI with NVLink**
- Breakthrough performance for GPU accelerated applications

**Asynchronous Checkpoint File Transfer**

**IBM HPC Software Solution**
- Parallel Performance Toolkit for Power
- Spectrum MPI
- ESSL / PESSL
- Spectrum Scale Platform LSF® Family
- xCAT
- XL, PGI & LLVM Compilers
- NVIDIA CUDA
- CSM
- Burst Buffer

**IBM HPC System Overview**
- New CORAL System Feature
- Collaboration between IBM and CORAL Labs
- Provides the Integration Framework for Next level System Management
- Open Source in 2018

**IBM Power Systems POWER9 server for HPC & AI with NVLink**
- Breakthrough performance for GPU accelerated applications

**Leverage Big Data capabilities to solve tough system management issues**

**IBM Log Analytics**

**IBM Spectrum LSF**

**xCAT (Open Source)**

**Leverage Big Data capabilities to solve tough system management issues**

**IBM Spectrum Scale**

**Burst Buffer**

**High Speed InfiniBand Network**
CORAL – IBM Delivers Summit and Sierra

- Deployments beginning with full acceptance in 2018
- Significant application performance over Titan (AMD/NVIDIA)
  - Achieved with ¼ the servers

3+EFLOPS
Tensor Ops

10X
Perf Over Titan

5-10X
Application Perf Over Titan
CORAL ORNL (Summit) 200PF System

**POWER9:**
- 22 Cores
- 4 Threads/core
- 0.54 DP TF/s
- 3.07 GHz

**Volta:**
- 7.0 DP TF/s
- 16GB @ 0.9 TB/s

**POWER9 2 Socket Server**
- 2 P9 + 6 Volta GPU
- 512 GiB SMP Memory (32 GiB DDR4 RDIMMs)
- 96 GiB GPU Memory (HBM stacks)
- 1.6 TB NVMe

**Mellanox IB4X EDR Switch IB-2**
- Mellanox IB4X EDR
- 648p Directors
- Full bisection

**ESS Building Block**
- SSC (4 ESS GL4): 8 servers, 16 JBOD
- 16.8 PB (gross)
- 38 KW max

**System**
- Standard 2U 19in. Rack mount Chassis
- Compute Rack
- 18 nodes
- 775 TF/s
- 10.7 TiB
- 59 KW max

**Floor plan rack concept**
- Compute (256)
- Switch (18)
- Storage (40)
- Infrastructure (4)
CORAL: ORNL Summit

Overhead H₂O distribution
CORAL LLNL (Sierra) 125PF System

POWER9: 2 Socket Server
- 2 P9 + 4 Volta GPU
- 256 GiB SMP Memory (16 GiB DDR4 RDIMMs)
- 64 GiB GPU Memory (HBM stacks)
- 1.6 TB NVMe

Volta:
- 7.0 DP TF/s
- 16GB @ 0.9 TB/s

POWER9: 22 Cores
- 4 Threads/core
- 0.54 DP TF/s
- 3.07 GHz

Volta:
- 7.0 DP TF/s
- 16GB @ 0.9 TB/s

Mellanox IB4X EDR Switch IB-2
- SXM2

Mellanox IB4X EDR
- 648p Directors
- Half Bisection

System

ESS Building Block
- SSC (4 ESS GL4):
  - 8 servers, 16 JBOD
  - 16.8 PB (gross)
  - 38 KW max

Compute Rack
- 18 nodes
- 523 TF/s
- 5.6 TiB
- 45 KW max

Floor plan rack concept
- Compute (240)
- Switch (9)
- Storage (24)
- Infrastructure (4)
CORAL: LLNL Sierra
Thank you!

ibm.com/systems/hpc