CONVERGENCE OF AI & HPC

THOR SEWELL
DIRECTOR OF WORKLOAD PRODUCT MARKETING, INTEL DATA CENTER GROUP
MARCH 12, 2018
AGENDA

• HPC
• Analytics
• Artificial Intelligence
• Convergence
• Customer Examples
• Exascale
THE HPC OPPORTUNITY

HPC MARKET
Est. HPC revenue ~$30B in 2021 at ~7% CAGR¹

DATA ANALYTICS
Server revenue CAGR ~17%; >$4 billion in 2021²

ARTIFICIAL INTELLIGENCE
DL in HPC server revenue CAGR ~81%, overall AI-HPC market ~30% CAGR by 2021³

CLOUD
>65% of HPC workloads are in private-hybrid clouds⁴

HPC WORKLOADS EXPANDING TO BECOME PERVERSIVE ACROSS INDUSTRIES

3 Source: Hyperion, Attributes of the Top National Buyers of HPC Resources for Deep Learning and Machine Learning, 2018
4 Source: Hyperion, Types of Clouds Used for HPC Workloads, 2017
The coming deluge of data
A treasure trove of valuable insights

Data → Insight

- Avg. internet user: 1.5 GB
- Autonomous vehicles: 4 TB
- Connected airplane: 5 TB
- Smart factory: 1 PB
- Cloud video providers: 750 PB

Source: Amalgamation of analyst data and Intel analysis.
DATA ANALYTICS EVOLUTION

Today

- Descriptive Analytics
- Diagnostic Analytics

Emerging

- Operational Analytics
- Predictive Analytics
- Prescriptive Analytics
- Cognitive Analytics
- Advanced Analytics

Self-Learning and Completely Automated Enterprise
Simulation-Driven Analysis and Decision-Making
Foresight

Insight
What Happened and Why
Hindsight
What Happened

DATA DELUGE
COMPUTE BREAKTHROUGH
INNOVATION SURGE
CONVERGENCE FOR GAINING INSIGHT

MODEL/SIMULATION

HIGH PERFORMANCE DATA ANALYTICS

ARTIFICIAL INTELLIGENCE

HIGH PERFORMANCE COMPUTING INFRASTRUCTURE
THREE PILLARS FOR CONVERGED ARCHITECTURE

Data
- One data lake
- High throughput
- Low latency
- Reduce data movement

Workload
- One cluster for diverse jobs
- Modeling & Sim
- Deep Learning
- Batch & real-time analytics
- Visualization

Platform HW&SW
- Software frameworks interoperability
- Cloud and data center
- Container & VM support
- Optimized
- Scalable

* Other names and brands may be claimed as the property of others.
Converging HPC and Big Data

Conceptual Architecture

Web Server nodes
Database nodes
Data Transfer nodes
Login nodes

Intel Omni-Path Architecture fabric

Parallel File System

ESM Nodes
12TB RAM
4 nodes

LSM Nodes
3TB RAM
42 nodes

RSM Nodes
128GB RAM
800 nodes,
48 with accel.

Intel Omni-Path
Architecture Case Study
· March 28, 2017

© 2017 Pittsburgh Supercomputing Center

* Other names and brands may be claimed as the property of others.
Fast & Efficient DL Scaling on CPU

- Convergence with Top1/5 > 74%/92%
- 97% scaling efficiency from 4 to 256 node runs.
  Batch size of 16 per node, Total TTT: 63 minutes


More Information
- IBM claims 95% scaling efficiency and Facebook claims 89%

Convergence with Top1/5 > 74%/92%
90% scaling efficiency. 4 - 256 node runs
Batch size of 32 per node, Total TTT: 70 Minutes
Deep Learning in Practice

Innovation Cycle
- Label data: 15%
- Load data: 15%
- Augment data: 23%
- Experiment with topologies: 15%
- Tune hyper-parameters: 15%
- Support inference inputs: 8%
- Document results: 8%

Time-to-Solution
- Source Data
- Development Cycle
- Inferencing
- Inference within broader application

Production Deployment
- Deploy
- Data Integration & Management
- Data Processing
- Decision Process
- Broader Application
- Refresh

Labor-intensive
- Compute-intensive (Training)

Support: Intel estimates
TECHNICAL CHALLENGES THAT MUST BE TACKLED TO ACHIEVE EXASCALE

• Massive Parallelism
• Memory and Storage
• Reliability
• Software
• Energy Consumption

ROADMAP GOALS MOVING FORWARD

Hasten Pace Of Architectural Innovation And Increase Cadence Of New Products

Deliver Improved, Real World Application Performance

Single Platform Scalable to Multiple Workloads – On-prem and in the Cloud
Notices & Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/benchmarks.

Benchmark results were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Cost reduction scenarios described are intended as examples of how a given Intel®-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

© 2018 Intel Corporation.

Intel, the Intel logo, and Intel Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as property of others.