Building Computer Architecture for the Era of AI and Multi-Cloud

H. Peter Hofstee, Ph.D.
IBM (& TU Delft)

March. 14, 2019

Infrastructure Matters...

when you deploy...

Hybrid multicloud

and tap into...

Cognitive workloads

A closer look at Summit & Sierra

#1 & #2 in HPC

… and > 3 ExaOp AI !

POWER9 hybrid multicloud

... IBM **Cloud** | vmware IBM **Cloud** Private NEW OPENSHIFT by Red Hat ...

Public cloud

Hybrid and private

IBM's Cognitive Systems portfolio for multicloud

Cloud-ready enterprise
& scale-out

Linux
clusters

Hyperconverged

Linux

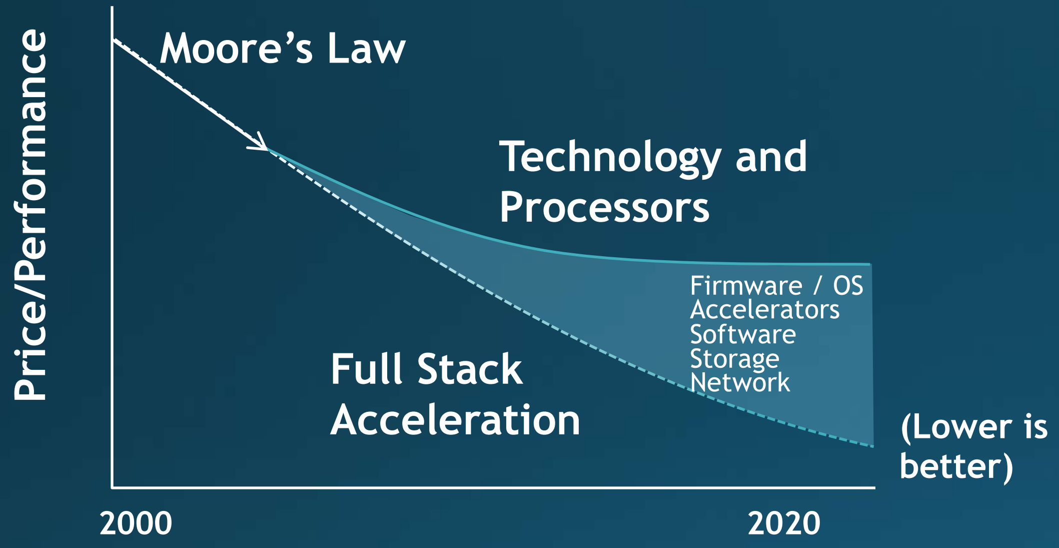
for Business

AIX

kubernetes

Fundamental forces are accelerating change in our industry

IT innovation can no longer come from just the processor



Full system stack innovation required

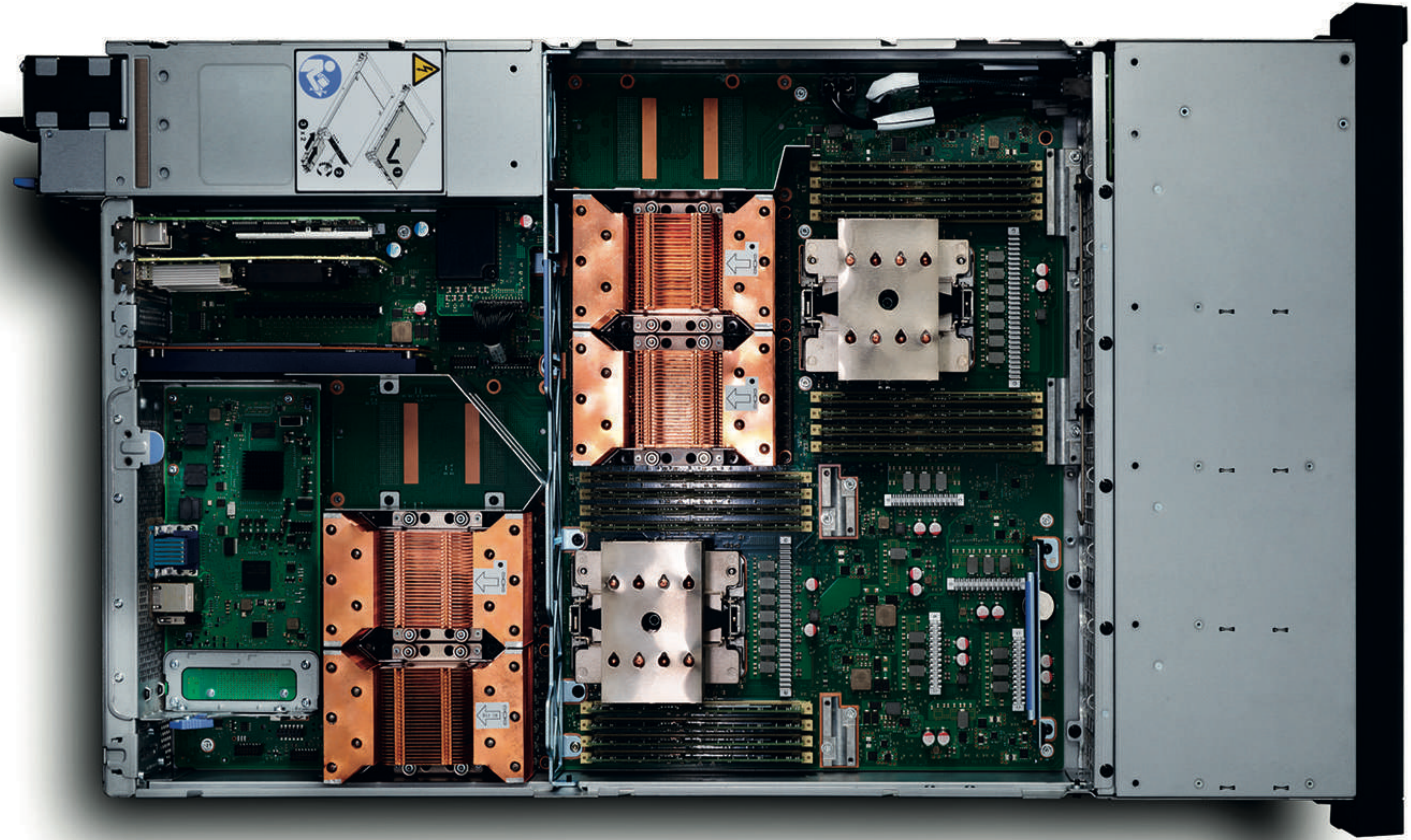
IT consumption models are expanding

Cognitive

Custom Hyperscale Data Centers

Hybrid Cloud

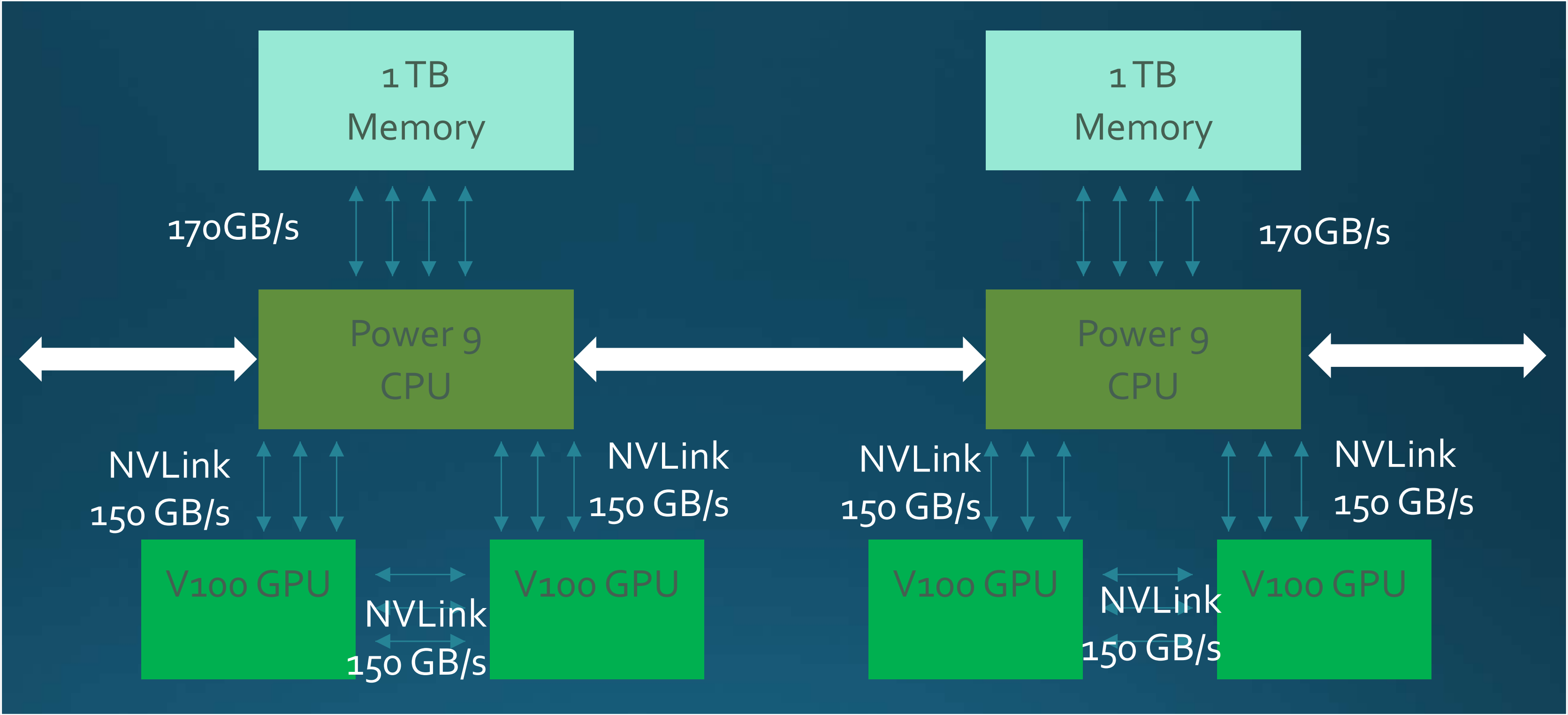
Open Solutions



5x Faster Data Communication with Unique CPU-GPU NVLink High-Speed Connection

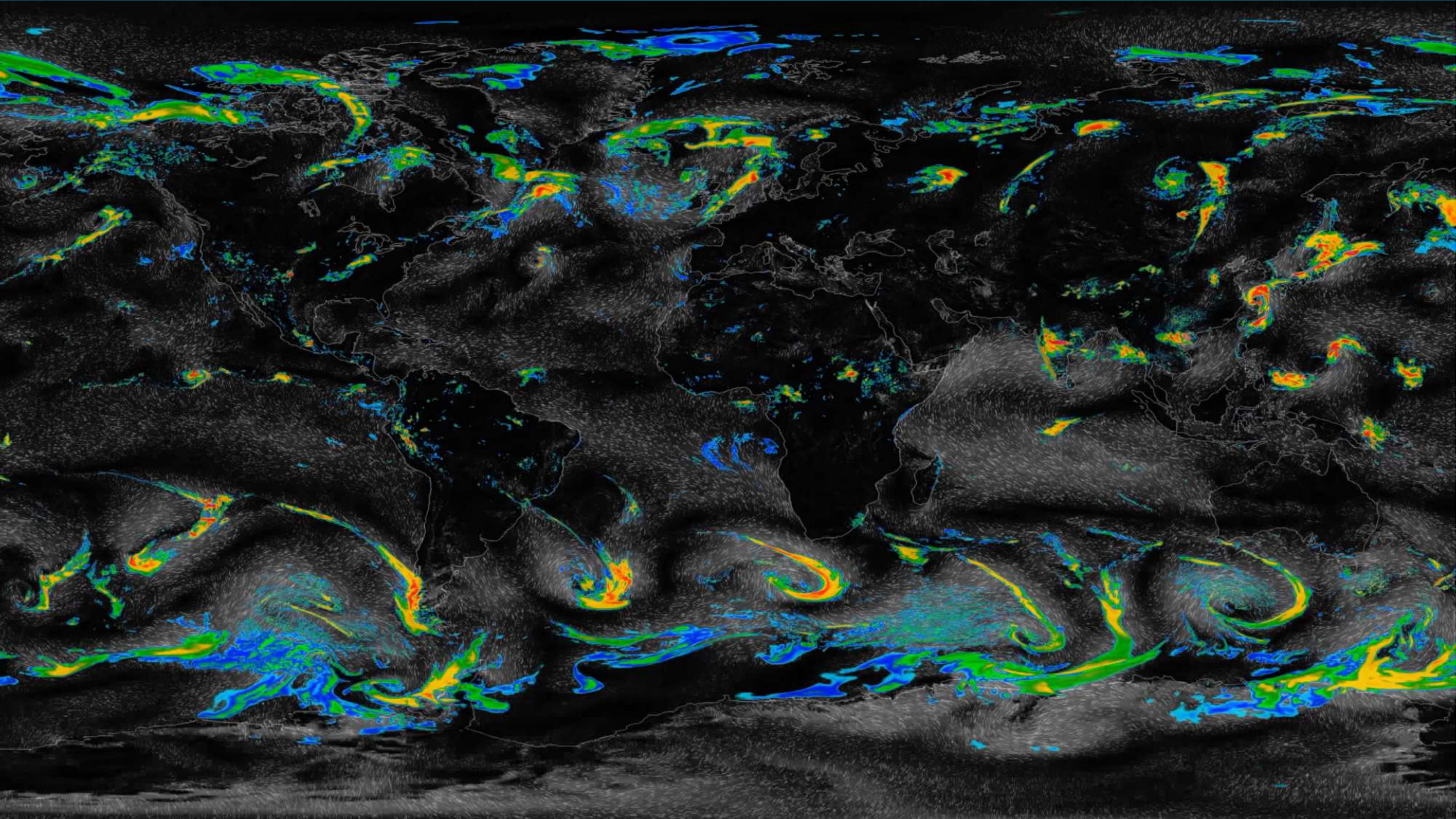
Store Large Models in System Memory

Fast Transfer via NVLink

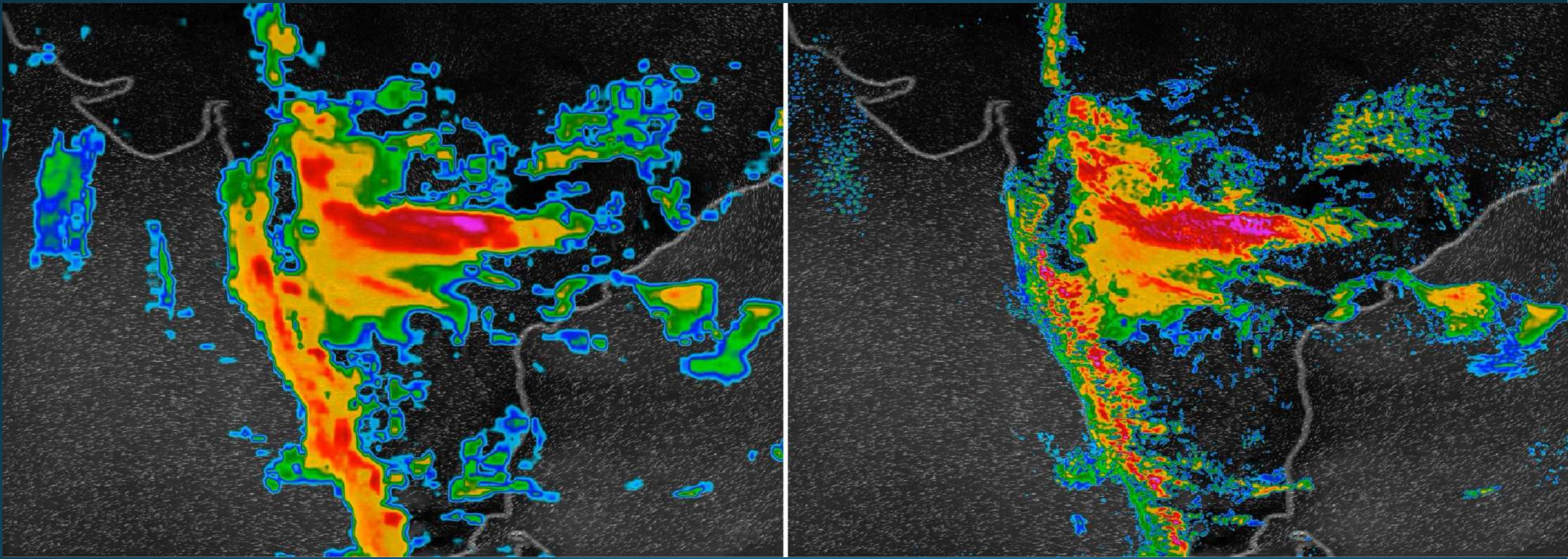


IBM AC922 Power System

Deep Learning Server (4-GPU Config)



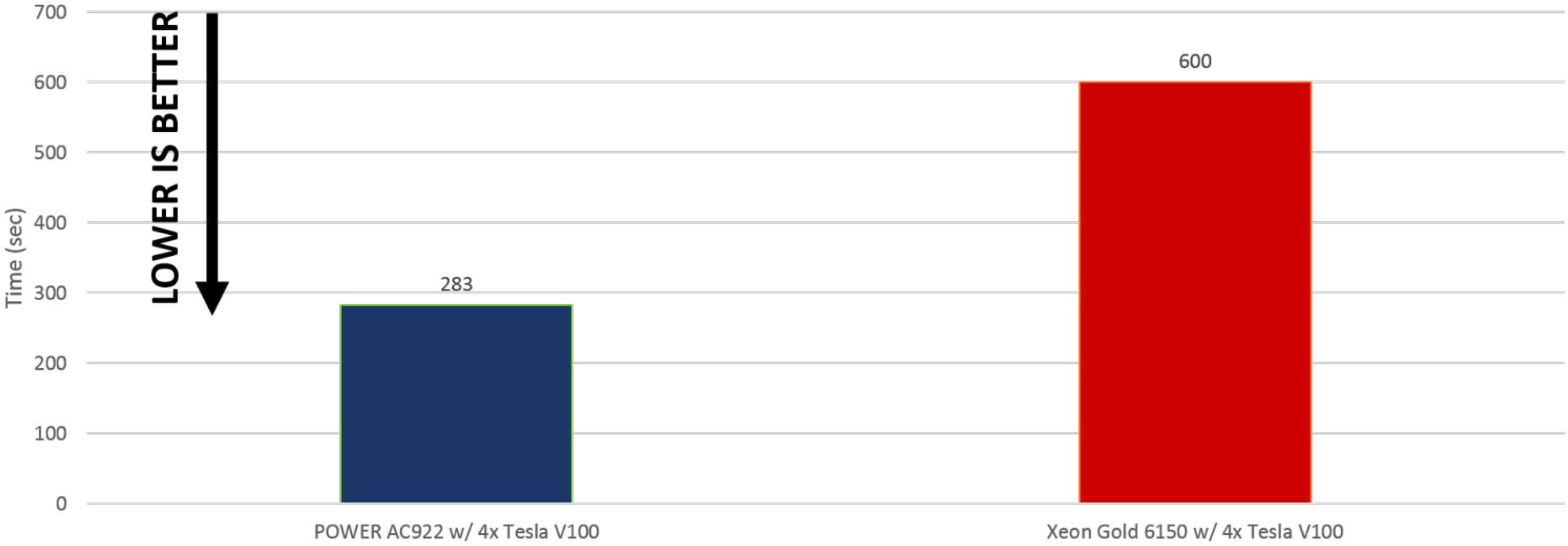
3km Resolution Weather Modeling on Power 9
84x AC922, Model Updates Every Hour



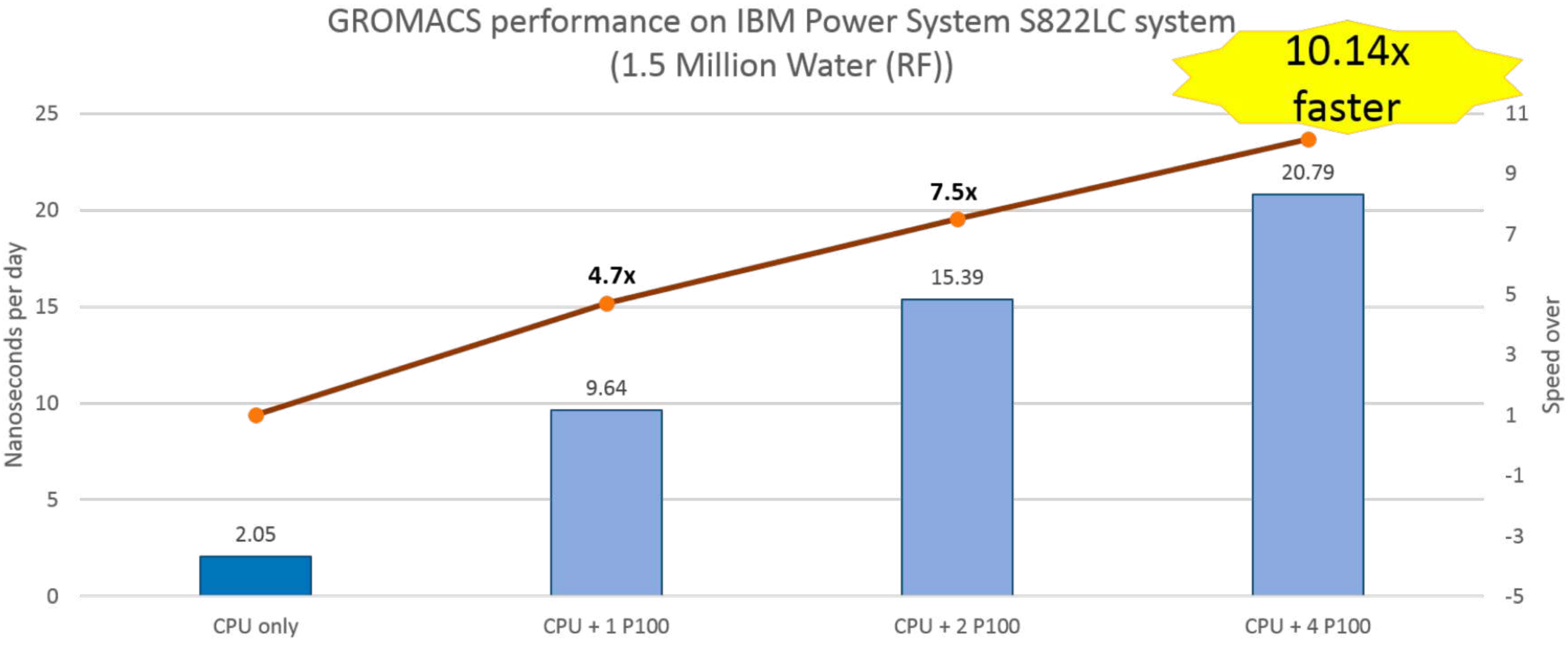
IBM Global High-Resolution Athmospheric Forecasting System

Molecular Dynamics (CPMD)

256 Water Random



https://developer.ibm.com/linuxonpower/perfcol/perfcol-technical/



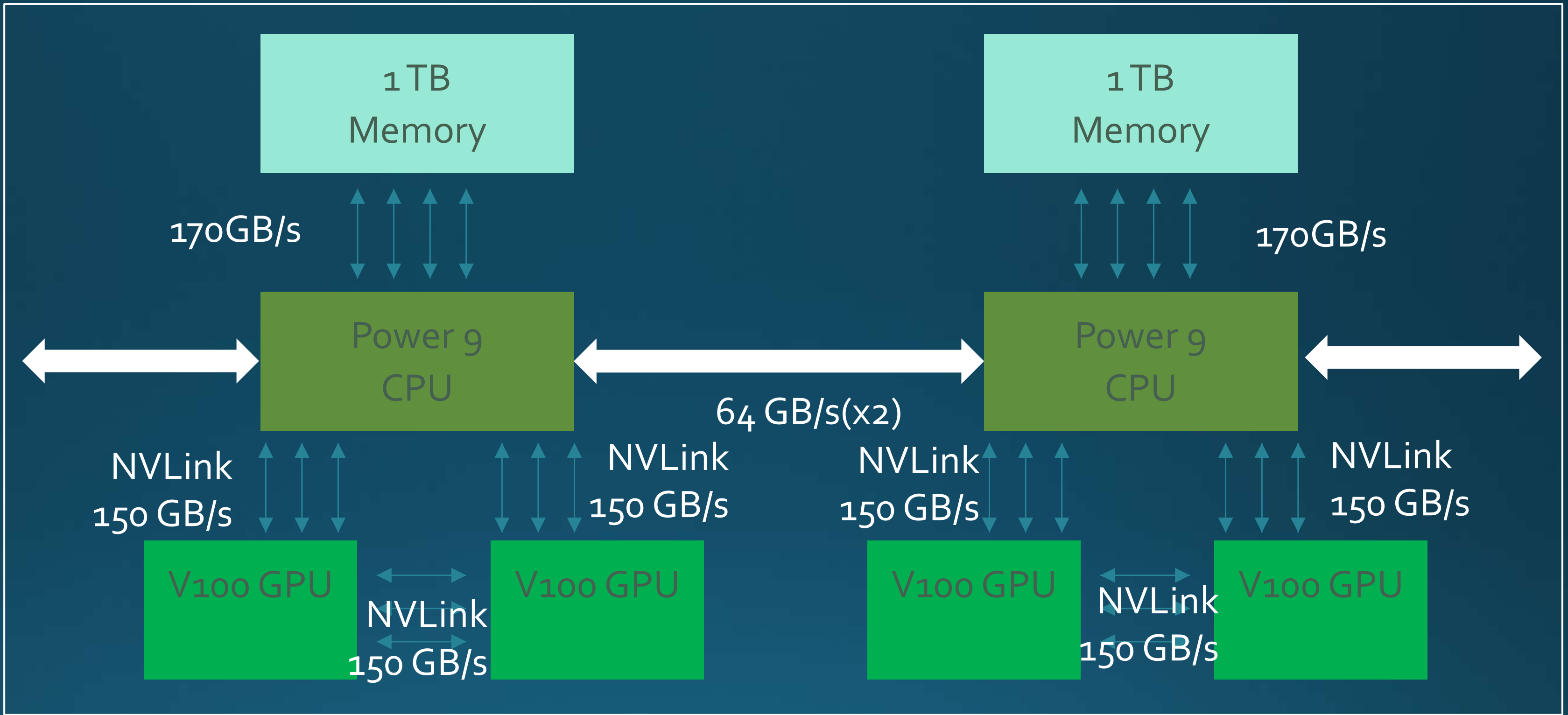
https://developer.ibm.com/linuxonpower/perfcol/perfcol-technical/

5x Faster Data Communication with Unique CPU-GPU NVLink High-Speed Connection

Store Large Models in System Memory

Fast Transfer via NVLink

Operate on One Layer at a Time



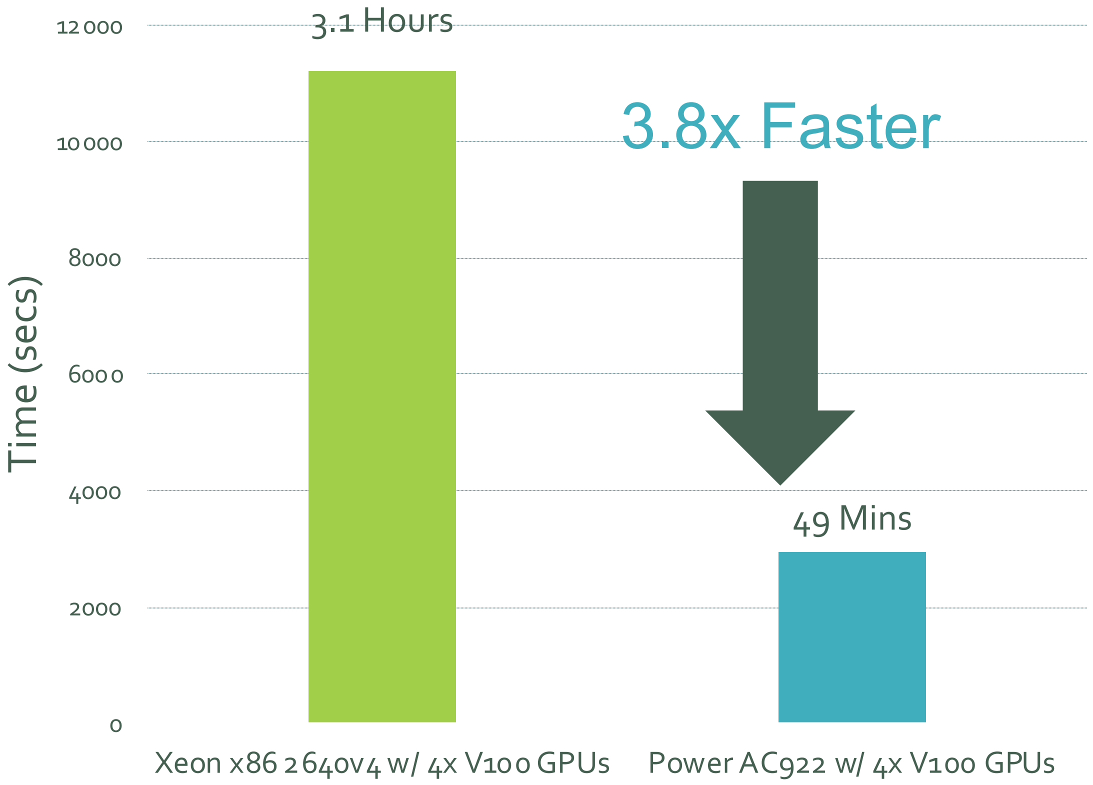
IBM AC922 Power System

Deep Learning Server (4-GPU Config)

Large AI Models Train ~4 Times Faster

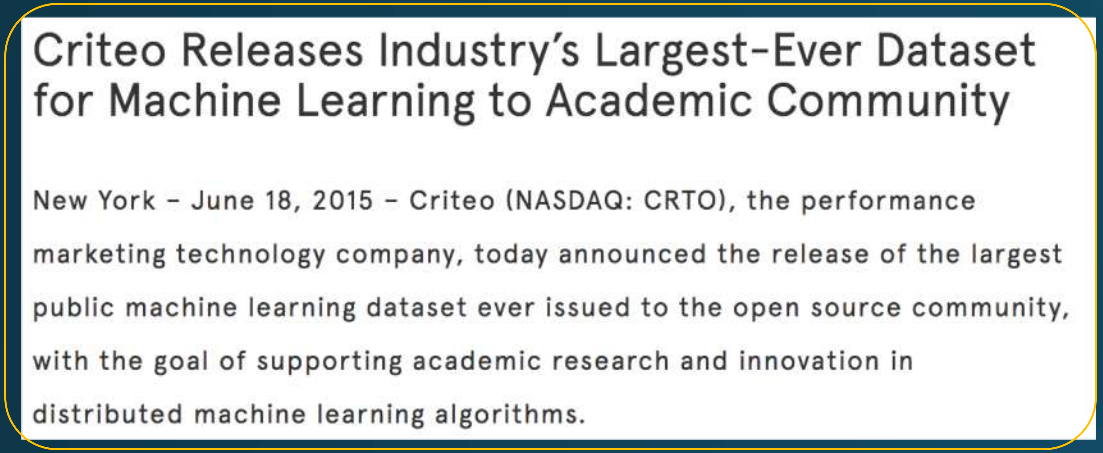
POWER9 Servers with NVLink to GPUs
vs
x86 Servers with PCIe to GPUs

Caffe with LMS (Large Model Support)
Runtime of 1000 Iterations



GoogleNet model on Enlarged ImageNet Dataset (2240x2240)

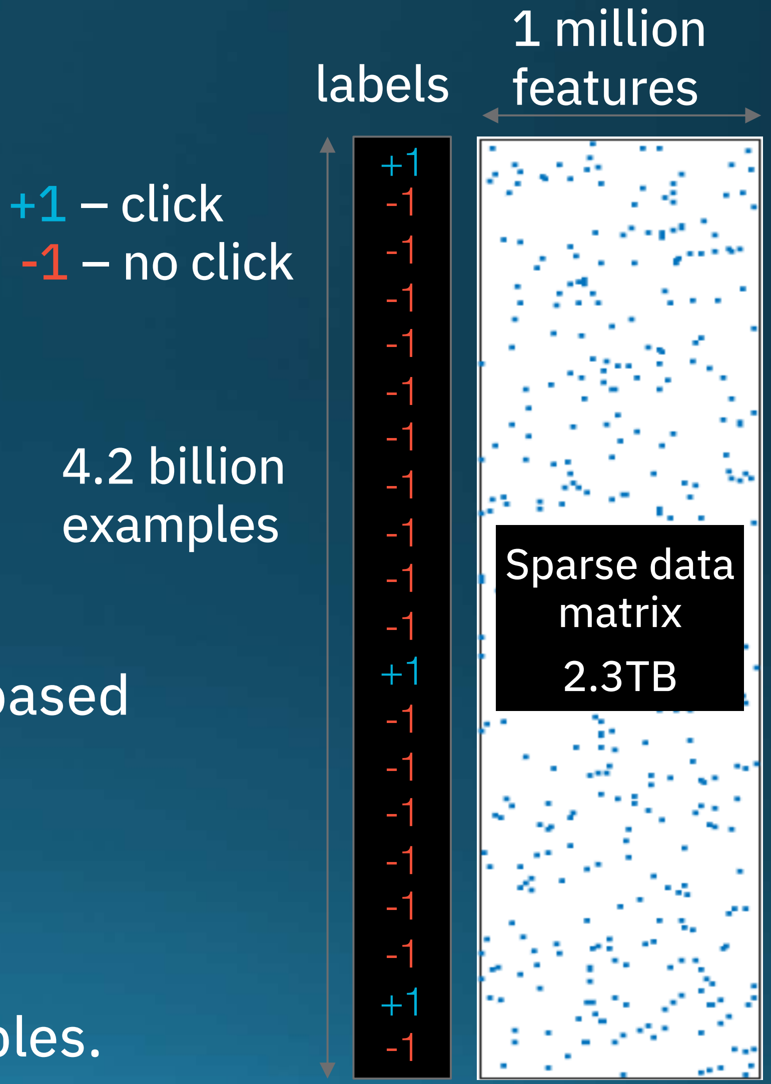
Tera-scale Computational Advertising Application



Criteo Releases Industry's Largest-Ever Dataset for Machine Learning to Academic Community

New York – June 18, 2015 – Criteo (NASDAQ: CRTO), the performance marketing technology company, today announced the release of the largest public machine learning dataset ever issued to the open source community, with the goal of supporting academic research and innovation in distributed machine learning algorithms.

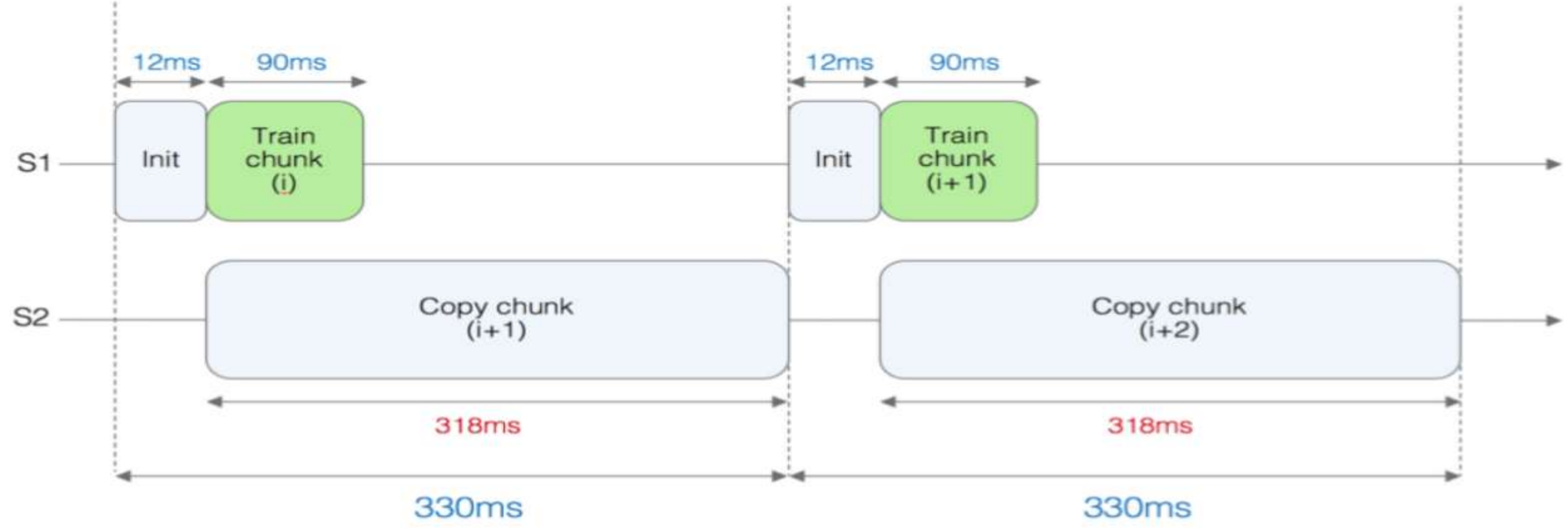
* Criteo Labs. 2015. Criteo Releases Industry□s Largest-Ever Dataset for Machine Learning to Academic Community. h□ps://www.criteo.com/news/press-releases/2015/07/criteo-releases-industrys-largest-ever-dataset/



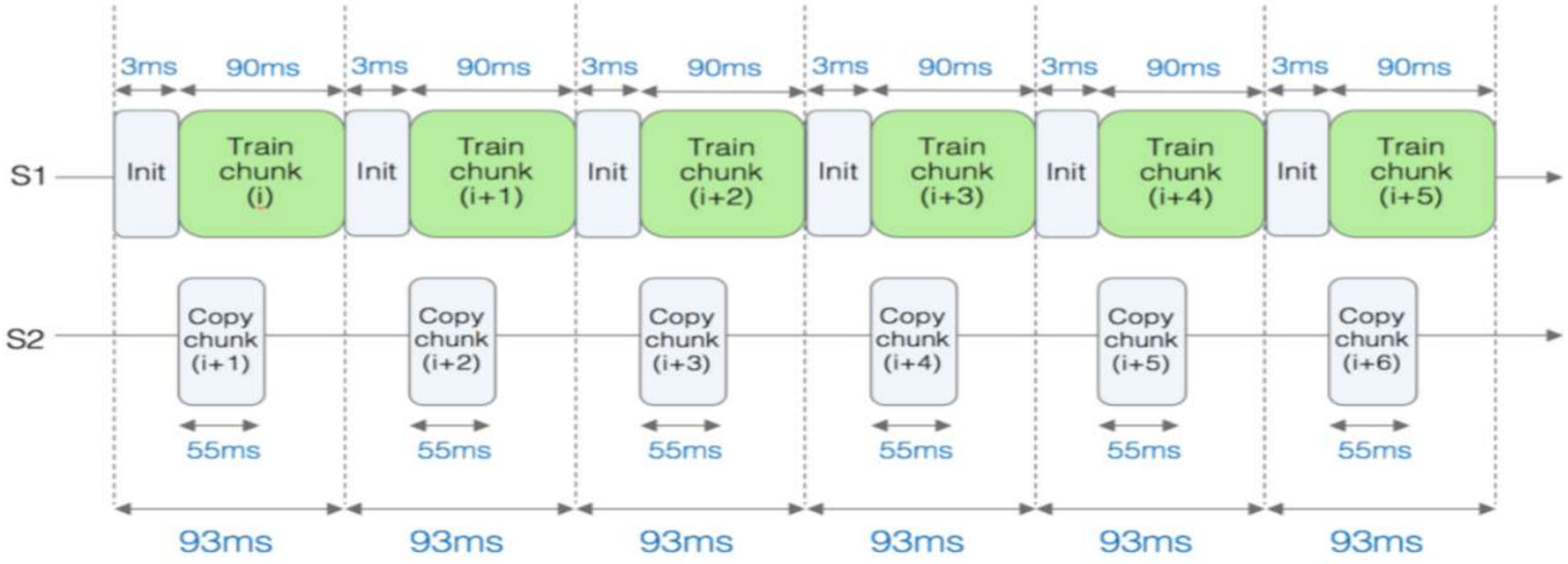
Goal: Predict whether a user will click on a given advert based on an anonymized set of features.

Train: Fit model parameters using **4.2 billion** examples.

Inference: Evaluate model on **180 million** unseen examples.



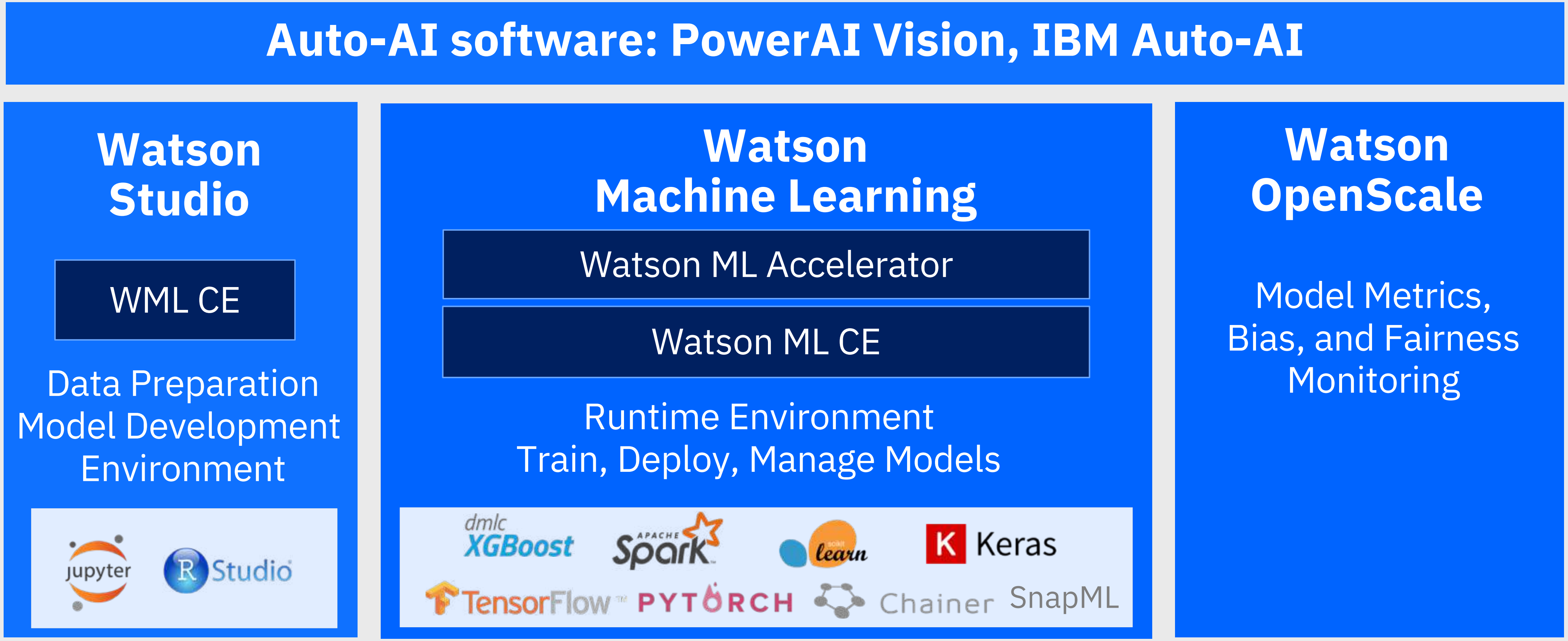
(a) Runtime profile on Intel x86 + PCIe Gen 3.0



(b) Runtime profile on POWER9 + NVLINK 2.0

Source IBM Research, Zurich

IBM Open Source Based AI Stack



Auto-AI software: PowerAI Vision, IBM Auto-AI

Watson Studio

WML CE

Data Preparation
Model Development
Environment

Watson Machine Learning

Watson ML Accelerator

Watson ML CE

Runtime Environment
Train, Deploy, Manage Models

Watson OpenScale

Model Metrics,
Bias, and Fairness
Monitoring

Accelerated AC922
Power9 Servers

Storage
(Spectrum Scale ESS)

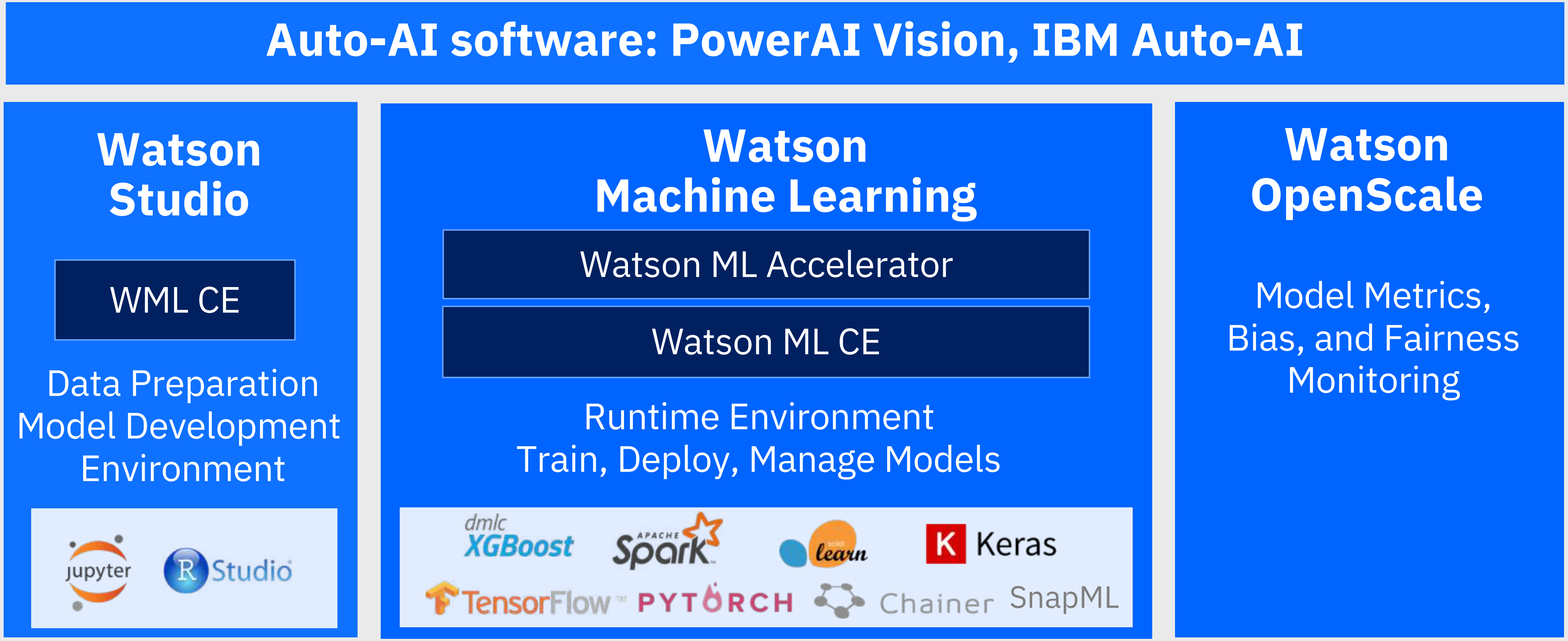
Previous Names:
WML Accelerator = PowerAI Enterprise
WML Community Ed. = PowerAI-base

Runs on x86 & other storage too

香港國際機場 | HONG KONG INTERNATIONAL AIRPORT

Hong Kong International Airport is leveraging POWER9 and PowerAI Vision to boost operational efficiency and improve security.

IBM Open Source Based AI Stack



Auto-AI software: PowerAI Vision, IBM Auto-AI

Watson Studio

WML CE

Data Preparation
Model Development
Environment

jupyter RStudio

Watson Machine Learning

Watson ML Accelerator

Watson ML CE

Runtime Environment
Train, Deploy, Manage Models

dmlc XGBoost, Apache Spark, scikit learn, Keras, TensorFlow, PYTORCH, Chainer, SnapML

Watson OpenScale

Model Metrics,
Bias, and Fairness
Monitoring

Accelerated AC922
Power9 Servers

Storage
(Spectrum Scale ESS)

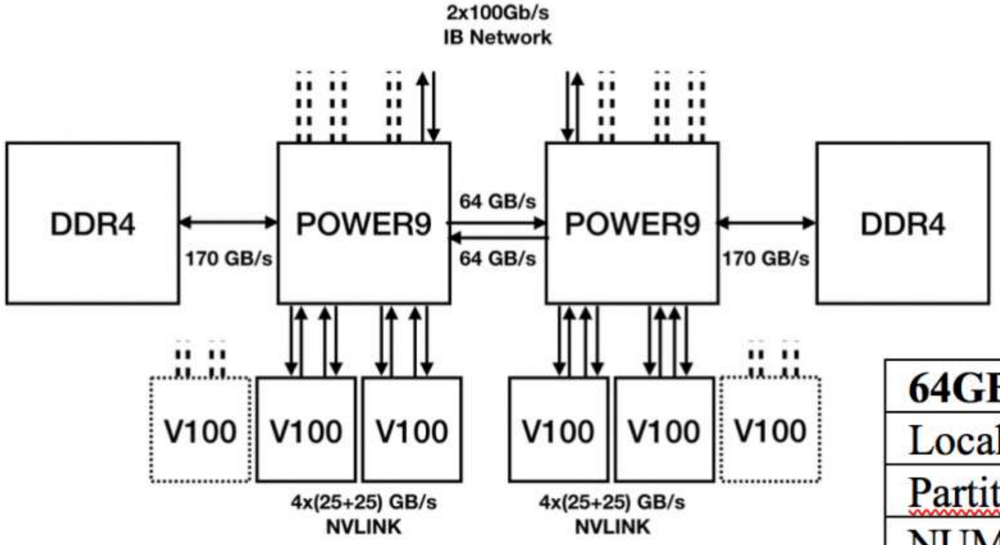
Previous Names:
WML Accelerator = PowerAI Enterprise
WML Community Ed. = PowerAI-base

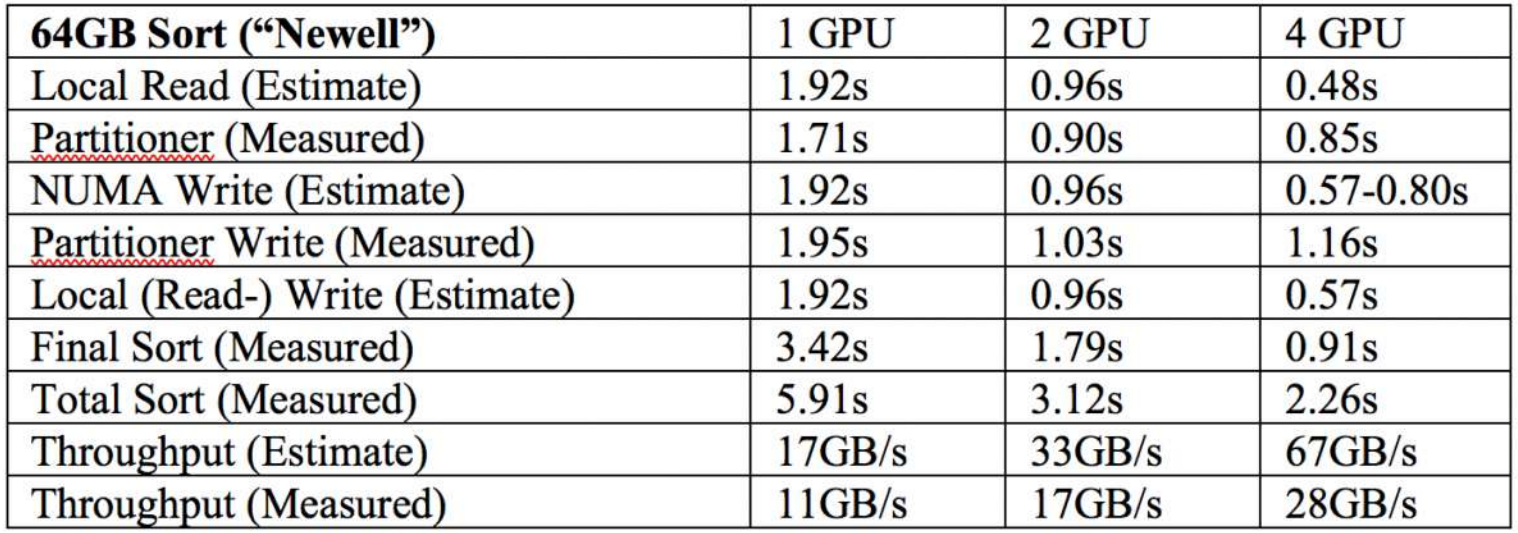
Runs on x86 & other storage too

A 64-GB Sort at 28 GB/s on a 4-GPU POWER9 Node for Uniformly-Distributed 16-Byte Records with 8-Byte Keys

Gordon C. Fossum[1], Ting Wang[2] and H. Peter Hofstee[1,3]

, Texas, USA
ghai, China
etherlands
m.com, hofstee@us.ibm.com





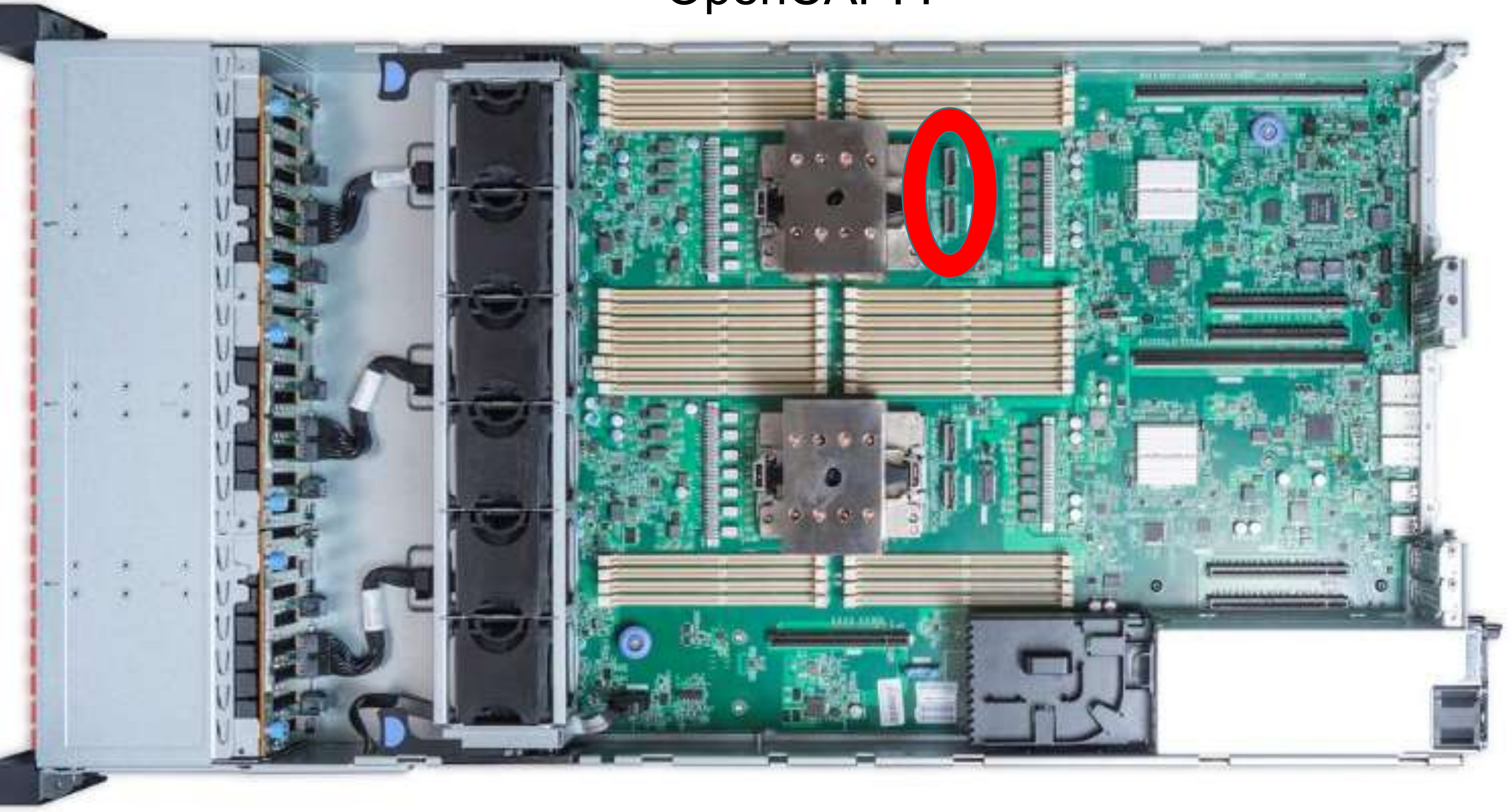
64GB Sort ("Newell")	1 GPU	2 GPU	4 GPU
Local Read (Estimate)	1.92s	0.96s	0.48s
Partitioner (Measured)	1.71s	0.90s	0.85s
NUMA Write (Estimate)	1.92s	0.96s	0.57-0.80s
Partitioner Write (Measured)	1.95s	1.03s	1.16s
Local (Read-) Write (Estimate)	1.92s	0.96s	0.57s
Final Sort (Measured)	3.42s	1.79s	0.91s
Total Sort (Measured)	5.91s	3.12s	2.26s
Throughput (Estimate)	17GB/s	33GB/s	67GB/s
Throughput (Measured)	11GB/s	17GB/s	28GB/s

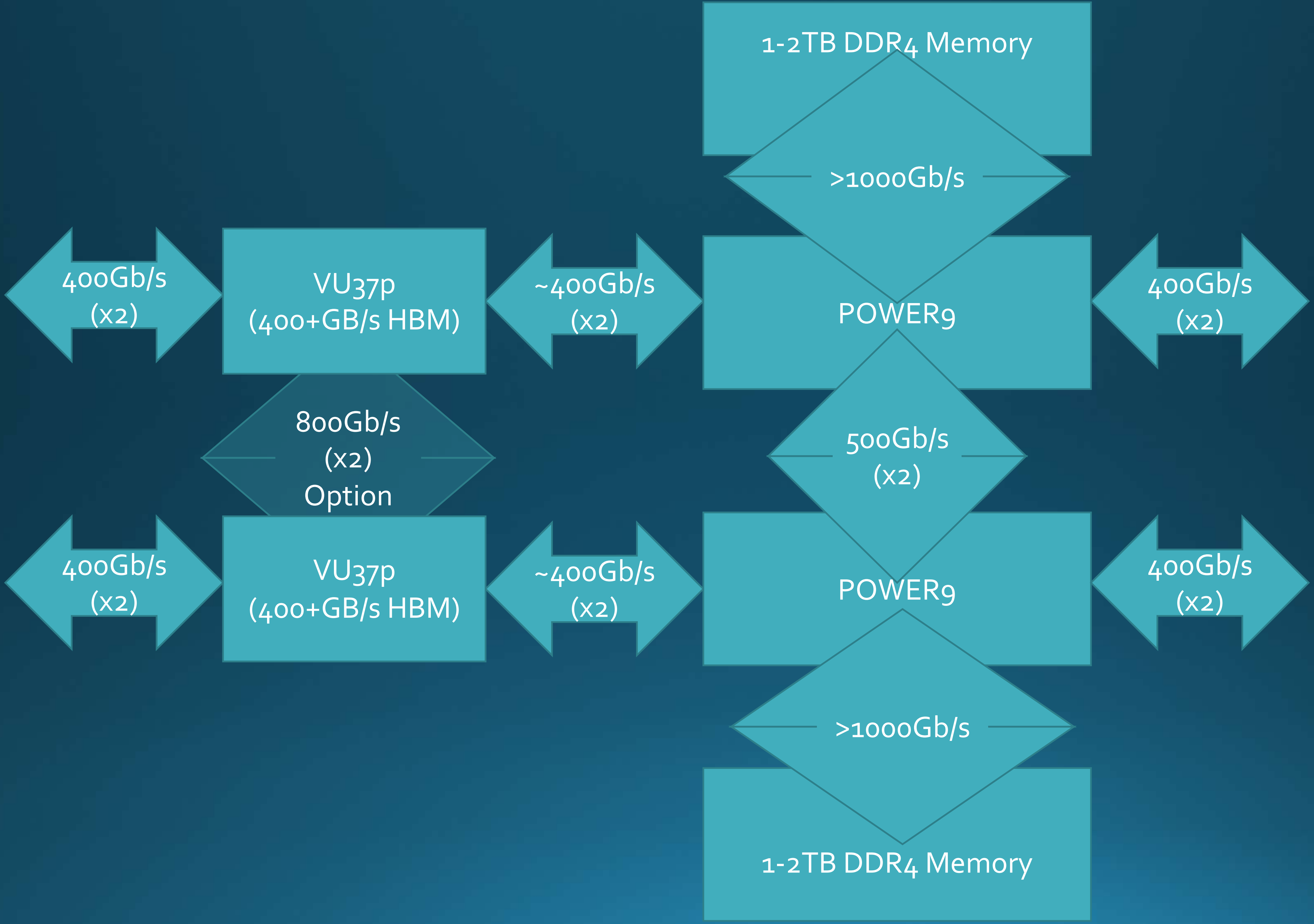
OpenPOWER Cloud-Optimized Systems

Power9 Zaius/Barreleye G2
1Tb/s (10x 100Gb/s) demo!

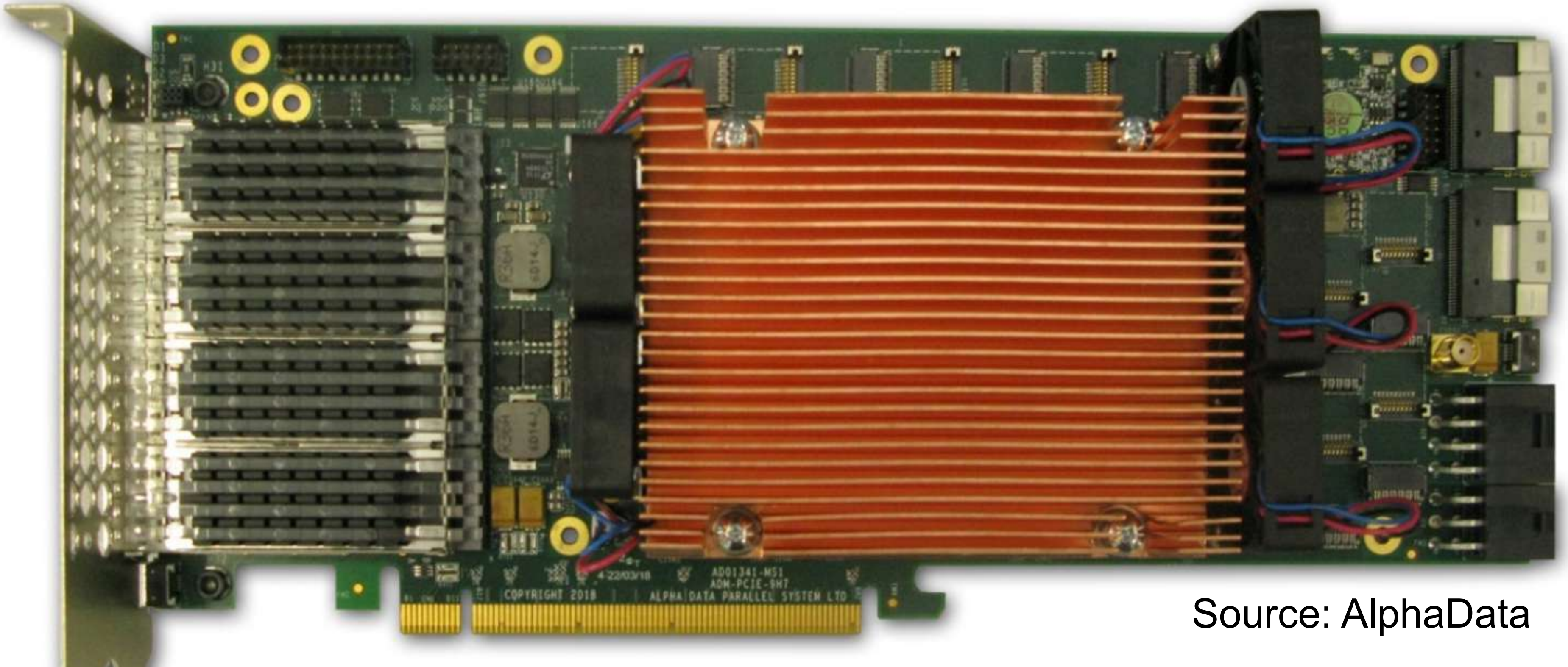
Wistron Power9 MiHawk

OpenCAPI !



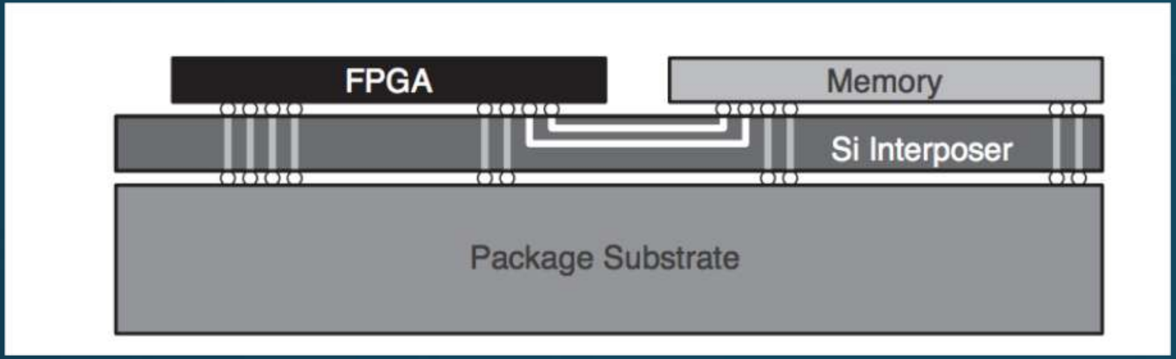

1-2TB DDR4 Memory
>1000Gb/s
400Gb/s (x2)
VU37p (400+GB/s HBM)
~400Gb/s (x2)
POWER9
400Gb/s (x2)
800Gb/s (x2) Option
500Gb/s (x2)
400Gb/s (x2)
VU37p (400+GB/s HBM)
~400Gb/s (x2)
POWER9
400Gb/s (x2)
>1000Gb/s
1-2TB DDR4 Memory

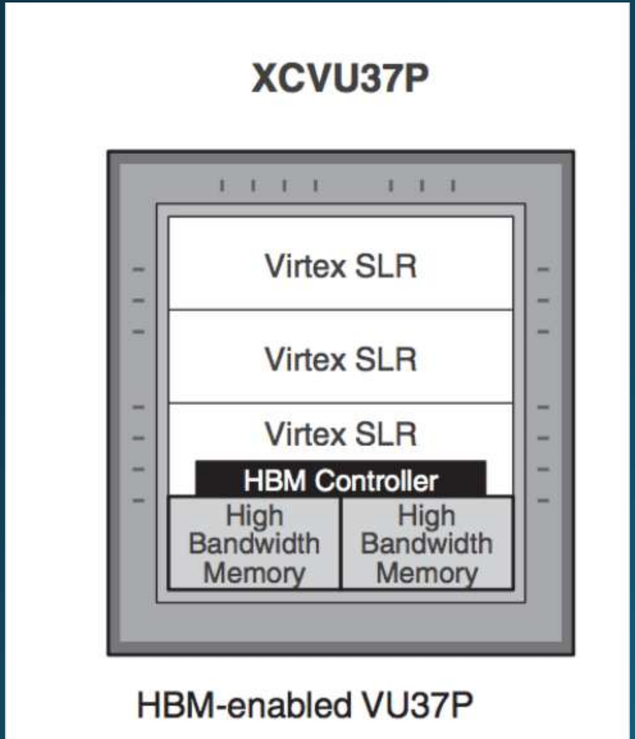
AlphaData ‘9H7 and ‘9V3 with OpenCAPI !



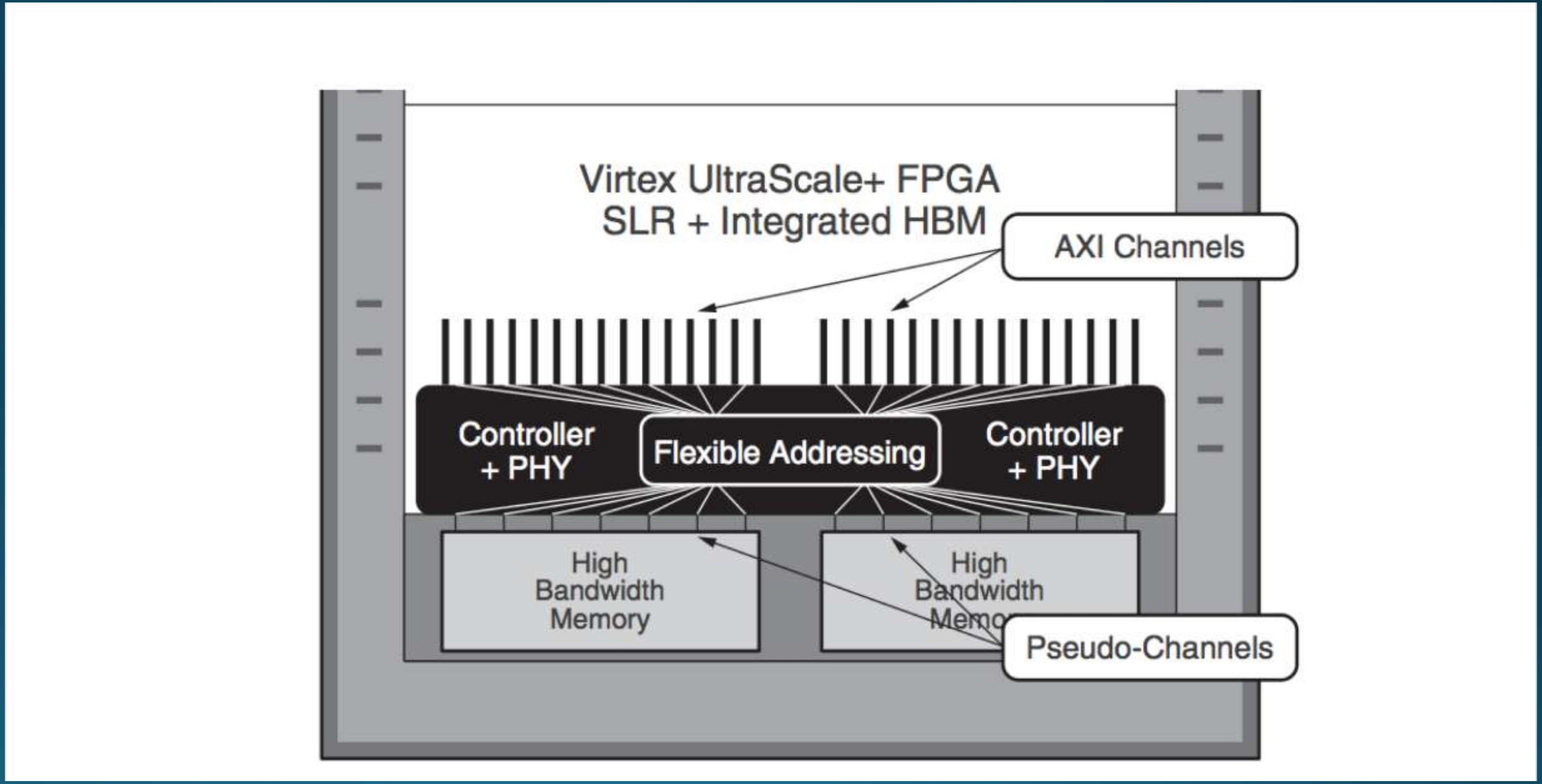
Source: AlphaData

Source: IBM





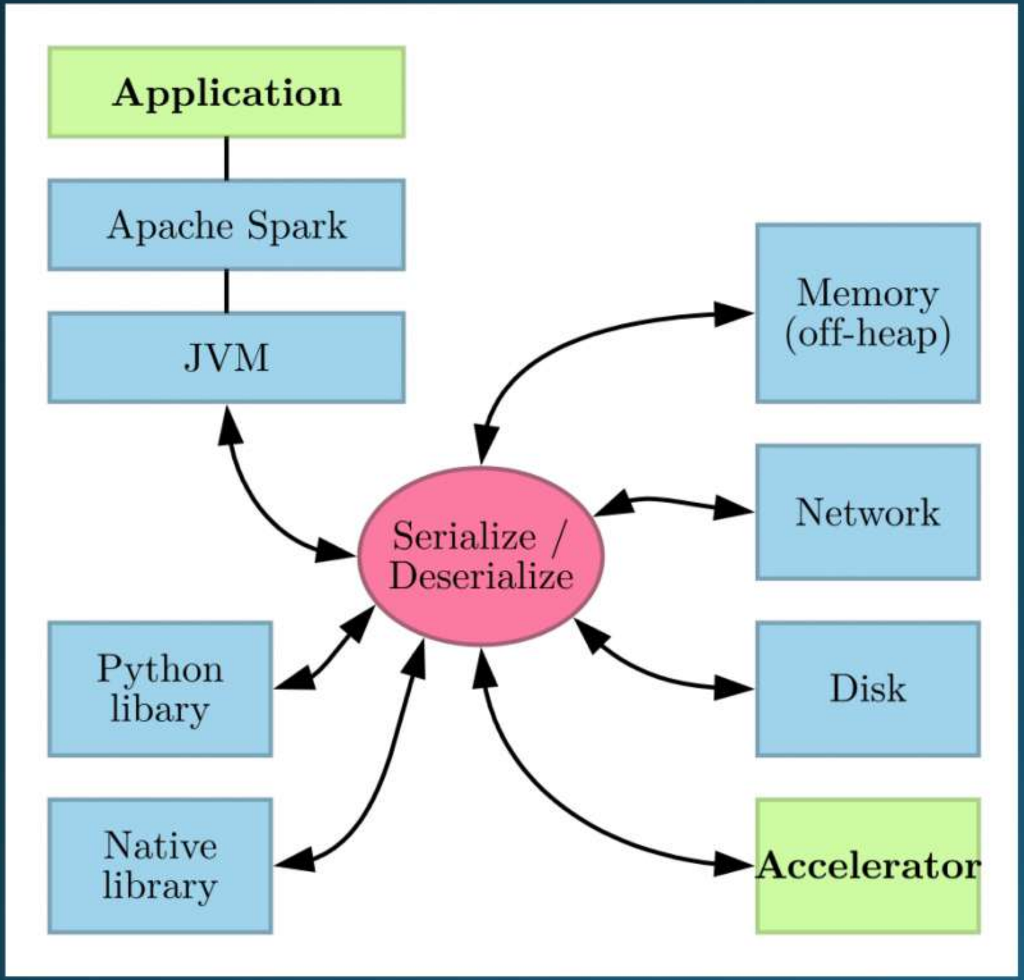
HBM-enabled VU37P

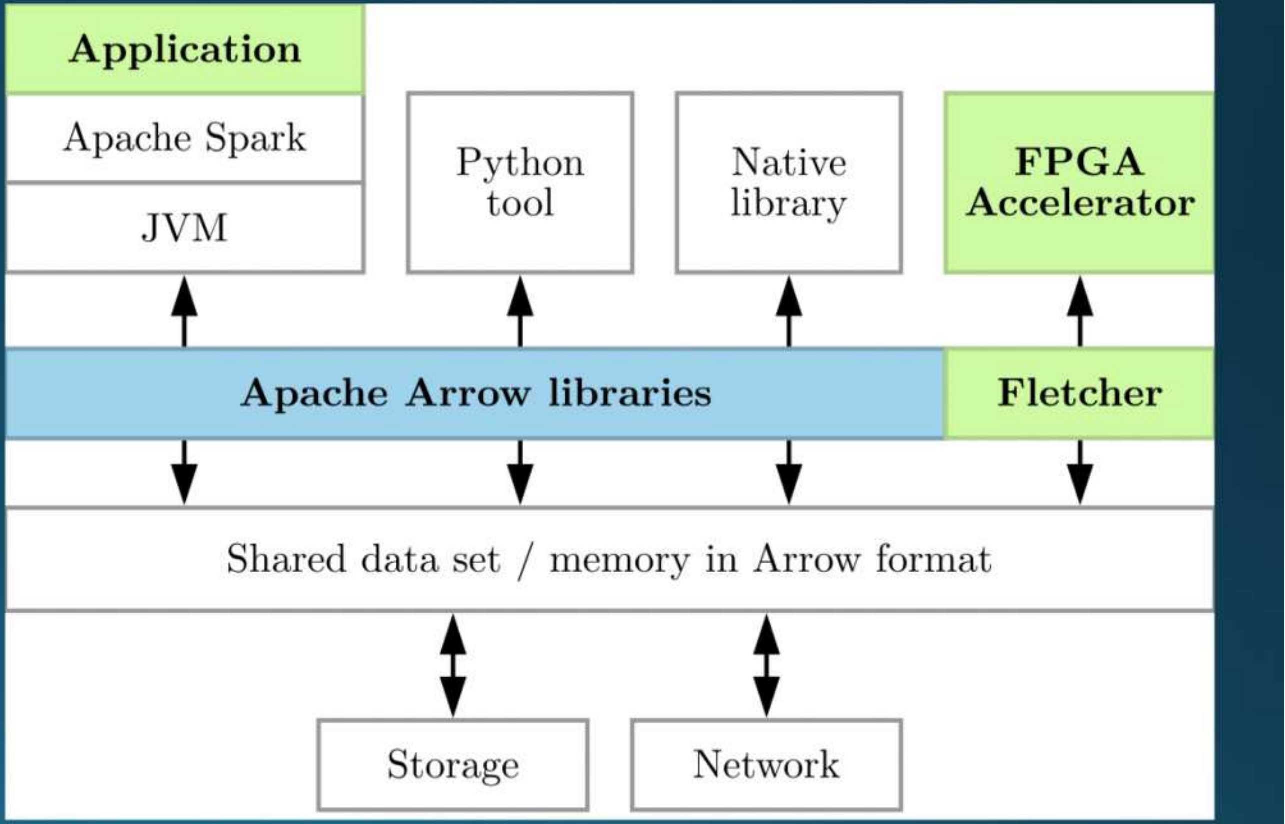


https://www.xilinx.com/support/documentation/white_papers/wp485-hbm.pdf

Old Way

Apache Arrow & Fletcher





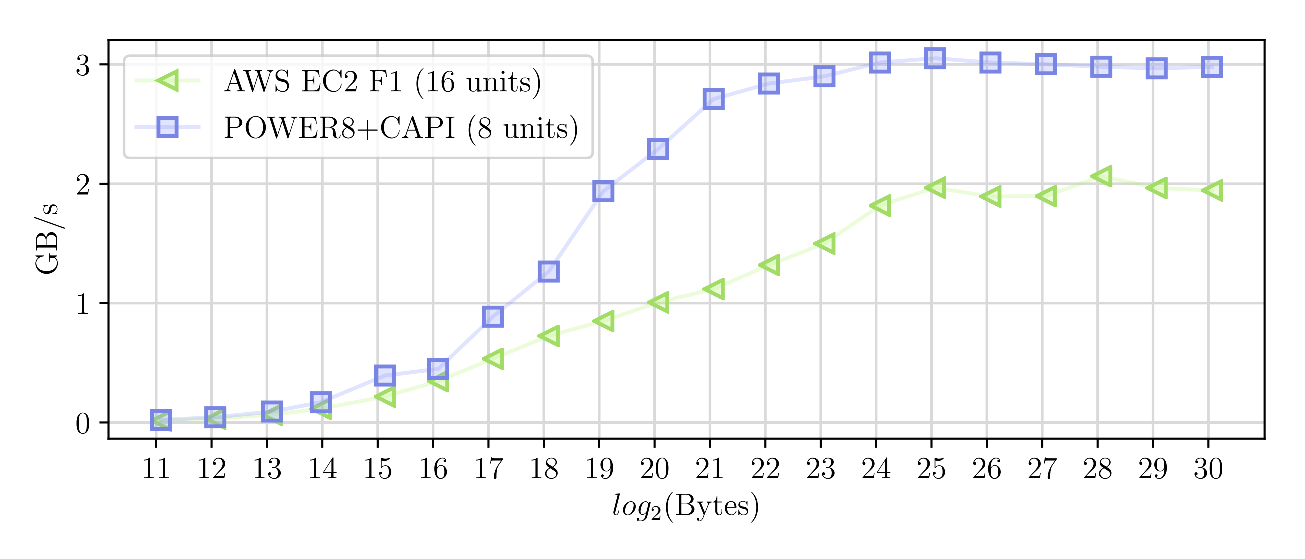
J. Peltenburg, e.a., TU Delft (OpenPOWER Summit USA 2018)

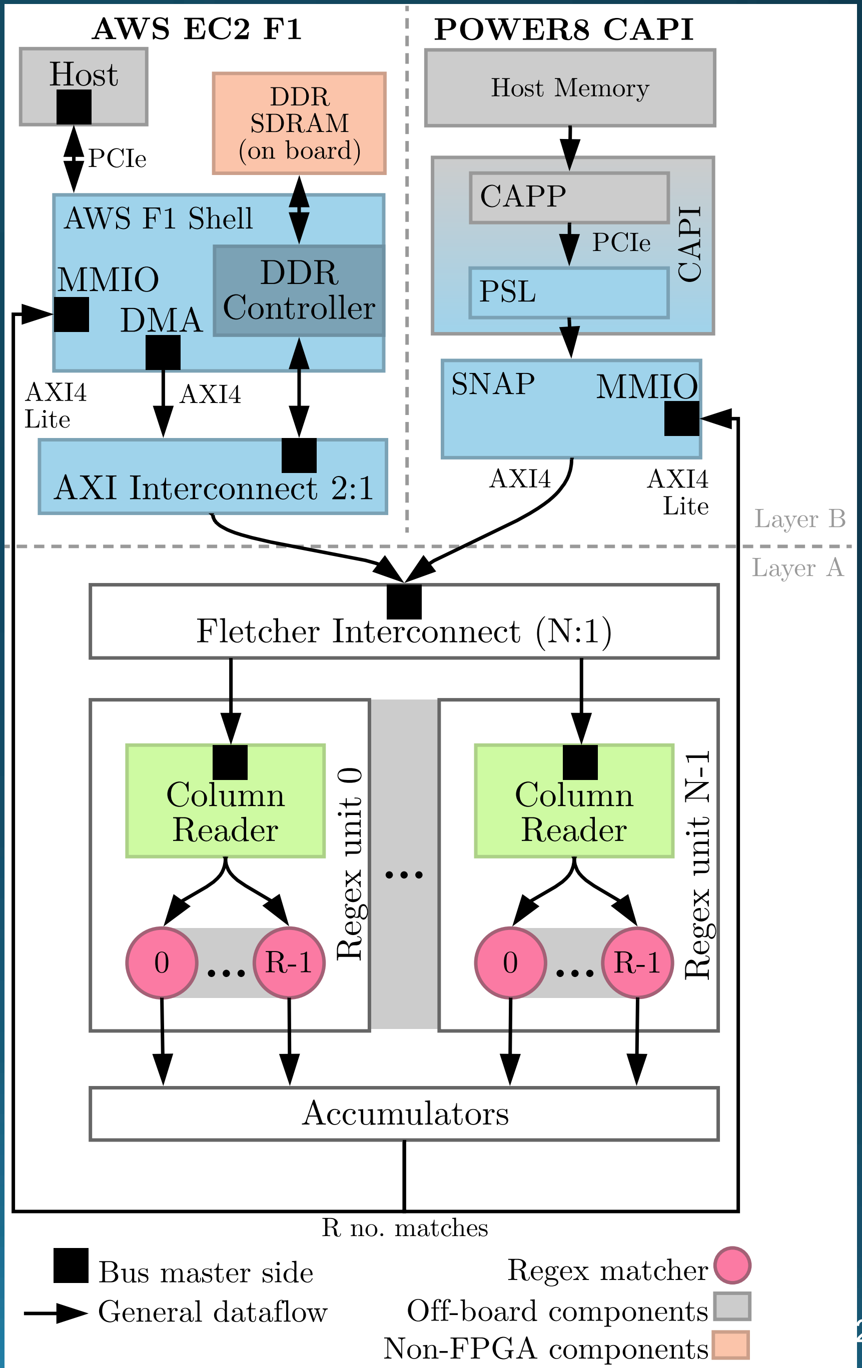
Regular expression matching

R=16 different regular expressions per unit

AWS EC2 F1:
- Virtex Ultrascale+
- N=16 regex units
- 256 regexes being matched in parallel

POWER8 CAPI (Supervessel, & soon at Nimbix):
- AlphaData KU3 (Kintex Ultrascale)
- N=8 regex units
- 128 regex being matched in parallel





Proposed POWER Processor Technology and I/O Roadmap

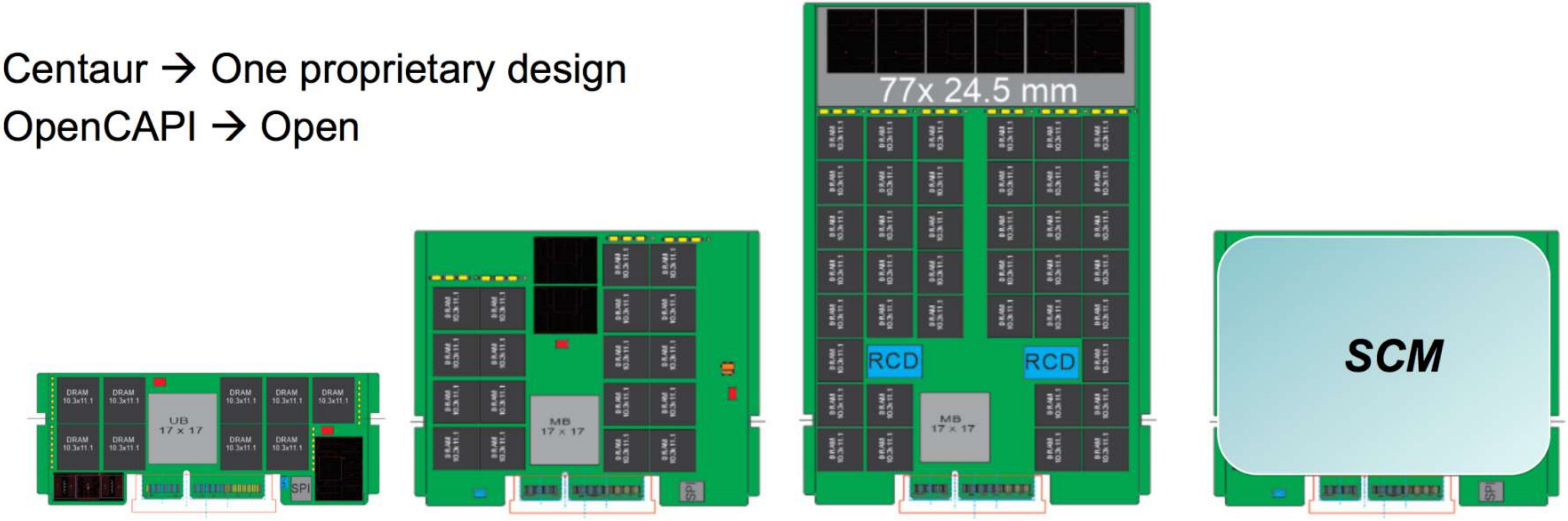
	POWER7 Architecture		POWER8 Architecture		POWER9 Architecture			POWER10
	2010 POWER7 8 cores 45nm New Micro-Architecture New Process Technology	2012 POWER7+ 8 cores 32nm Enhanced Micro-Architecture New Process Technology	2014 POWER8 12 cores 22nm New Micro-Architecture New Process Technology	2016 POWER8 w/ NVLink 12 cores 22nm Enhanced Micro-Architecture With NVLink	2017 P9 SO 24 cores 14nm New Micro-Architecture Direct attach memory New Process Technology	2018 P9 SU 24 cores 14nm Enhanced Micro-Architecture Buffered Memory	2019 P9 w/ Adv. I/O 24 cores 14nm Enhanced Micro-Architecture New Memory Subsystem	2020+ P10 TBD cores New Micro-Architecture New Technology
Sustained Memory Bandwidth	Up To 65 GB/s	Up To 65 GB/s	Up To 210 GB/s	Up To 210 GB/s	Up To 150 GB/s	Up To 210 GB/s	Up To 350 GB/s	Up To 435 GB/s
Standard I/O Interconnect	PCIe Gen2	PCIe Gen2	PCIe Gen3	PCIe Gen3	PCIe Gen4 x48	PCIe Gen4 x48	PCIe Gen4 x48	PCIe Gen5
Advanced I/O Signaling	N/A	N/A	N/A	20 GT/s 160GB/s	25 GT/s 300GB/s	25 GT/s 300GB/s	25 GT/s 300GB/s	32 & 50 GT/s
Advanced I/O Architecture	N/A	N/A	CAPI 1.0	CAPI 1.0 , NVLink 1.0	CAPI 2.0, OpenCAPI3.0, NVLink2.0	CAPI 2.0, OpenCAPI3.0, NVLink2.0	CAPI 2.0, OpenCAPI4.0 NVLink3.0	TBD

Statement of Direction, Subject to Change

OpenCAPI Memory

- Signaling → AXON @25.6GHz vs DDR4 @ 3200 MHz
  - 4x bw per signal IO
- Idle latency over traditional DDR
  - POWER8/9 Centaur design ~10 ns
  - OpenCAPI target of ~5 ns

- Centaur → One proprietary design
- OpenCAPI → Open



J. Stuecheli, IBM: OpenPOWER Summit Europe 2018

Conclusions

- It’s about more than the CPU cores
  - Even though POWER9 cores are very good too!
- Investment in IO & OpenPOWER collaborations pays off
  - Better acceleration – better BW, lower latency, better CPU utilization with GPU & FPGA
  - Better networking – better BW, lower latency, lower CPU
  - Better memory/storage – better BW, lower latency, lower CPU
- Use examples:
  - HPC – Coral systems
  - Big Data – sort (10x per node of current sortbenchmark.org leader)
  - AI – large models (3.5-4x faster on large models)
- Open Hardware – Open Standards – Based on Open Software:
  - Multicloud

And a call to arms …

- Lots of opportunities for research & collaboration
  - Changing system architecture landscape

- Many OpenPOWER systems available from many vendors
  - Open ecosystem
  - Open firmware (leveraged e.g. by Talos Raptor systems for a more secure workstation)
  - Shared memory accelerator architecture

- Besides high-BW GPU many exciting new opportunities with FPGAs
  - Interface new memory types with OpenCAPI 3.0/3.1
  - Extreme network bandwidth
  - HBM
  - Near-storage computing (e.g. CAPI-attached flash or SCM)

Legal notices

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectually property rights, may be used instead.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER OR IMPLIED. IBM LY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. IBM makes no representations or warranties, ed or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 1 0504- 785
U.S.A.

Information and trademarks

IBM, the IBM logo, ibm.com, IBM System Storage, IBM Spectrum Storage, IBM Spectrum Control, IBM Spectrum Protect, IBM Spectrum Archive, IBM Spectrum Virtualize, IBM Spectrum Scale, IBM Spectrum Accelerate, Softlayer, and XIV are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at http://www.ibm.com/legal/copytrade.shtml

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

ITIL is a Registered Trade Mark of AXELOS Limited.

UNIX is a registered trademark of The Open Group in the United States and other countries.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This presentation and the claims outlined in it were reviewed for compliance with US law. Adaptations of these claims for use in other geographies must be reviewed by the local country counsel for compliance with local laws.

Special notices

This document was developed for IBM offerings in the United States as of the date of publication. IBM may not make these offerings available in other countries, and the information is subject to change without notice. Consult your local IBM business contact for information on the IBM offerings available in your area.

Information in this document concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The information contained in this document has not been submitted to any formal IBM test and is provided "AS IS" with no warranties or guarantees either expressed or implied.

All examples cited or described in this document are presented as illustrations of the manner in which some IBM products can be used and the results that may be achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions.

IBM Global Financing offerings are provided through IBM Credit Corporation in the United States and other IBM subsidiaries and divisions worldwide to qualified commercial and government clients. Rates are based on a client's credit rating, financing terms, offering type, equipment type and options, and may vary by country. Other restrictions may apply. Rates and offerings are subject to change, extension or withdrawal without notice.

IBM is not responsible for printing errors in this document that result in pricing or information inaccuracies.

All prices shown are IBM's United States suggested list prices and are subject to change without notice; reseller prices may vary.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this document may have been estimated through extrapolation. Users of this document should verify the applicable data for their specific environment.