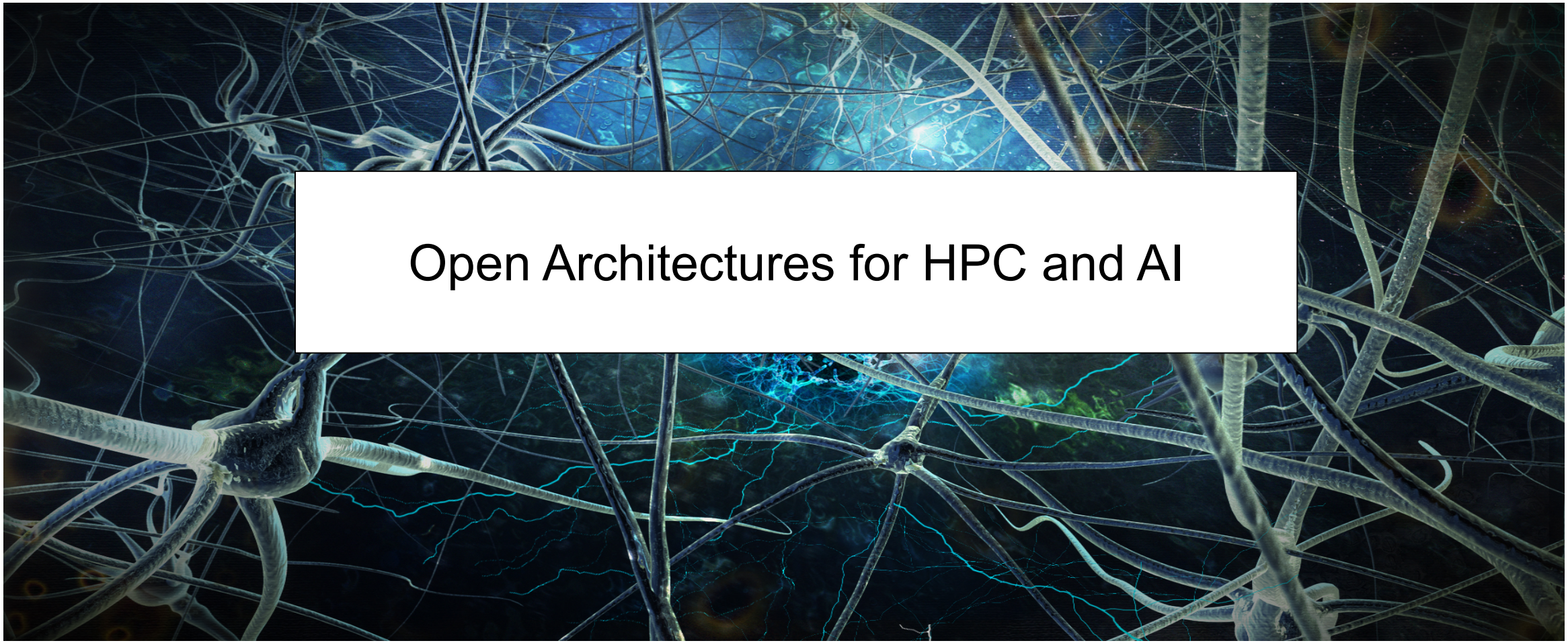


Open Architectures for HPC and AI

Ander Ochoa Gilo – ander.ochoa.gilo@ibm.com
Cognitive Systems Technical Architect for SPGI
OpenPOWER Foundation member

https://es.linkedin.com/in/anderotxoa

@AnderOtxoa

Current Mayor Architectures

PROPIETARY ARCHITECTURES

X86
- Intel
- AMD

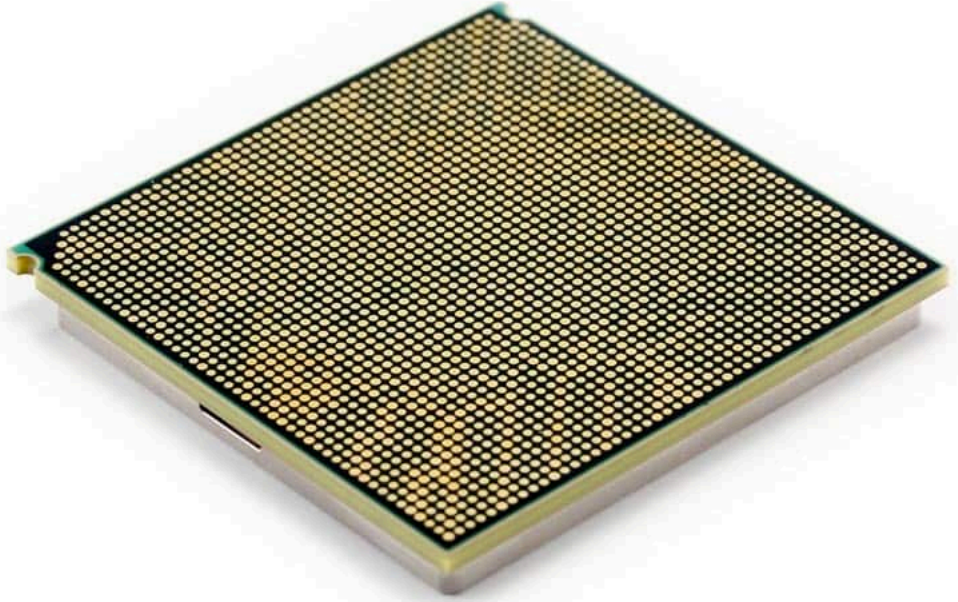
LICENCEABLE ARCHITECTURE

ARM
- Samsung
- Apple
- …

OPEN ARCHITECTURES

- RISC-V
- **OpenPOWER**

POWER9



OpenPOWER Foundation (members @ 09/2019)

Platinum

Google

IBM

inspur 浪潮

YADRO

Gold

HITACHI Inspire the Next

redhat.

wistron

Silver

2CRSI

archanan

ADDFOR s.p.a.

Aerobyte

Bowmicro, Ltd.

AEWIN

ALPHA DATA

amwin

Altos Computing, Inc.

appentra

AXYON.AI

7Elements LLC

Avery Design Systems

Beijing Auphi Bi Co., Ltd.

Beijing Innovation Open Source Technology Co. Ltd.

Beijing Rungian Systems Co., Ltd.

Beijin Sunway XinYue Technology Company Limited

YAN IRONG

bioretics

BITS

Bitfusion

BlazingDB, Inc.

BLUEBEE

boreo

BOSTON

BROADCOM

Bright Computing, Inc.

Brytlyt

Bull

C8 Technology Group (Pty.) Ltd.

Celestica

Chelsio Communications

WwIT

Chinaoly Technology Co. Ltd.

CoCoLink

Datalog, Inc.

Deep Cognition

D.F.I

Dolphin

E4 COMPUTER ENGINEERING

EACT Technologies PTE LTD

Eidetic Communications, Inc.

ellexus

Emmion

eoxys

EVERSPIN

FIXSTARS

GRAID TECH

T2i

H3

FLYSLICE

Shanghai ThinkForce Electronic Technology Co., Ltd.

Shenyang China-Bigdata Technology Co., Ltd.

Shenyang Yanyun Cloud Computing Technology Co., Ltd.

DaoCloud

Shanghai Weida Cloud Power Co., Ltd.

Shenzhen Yu-Yuan Intelligence

SK hynix

skymind

Smart IQ Robots, LLC

Smart Modular Technologies

SME Up Spa

Solarflare Communications

Splitted-Desktop Systems

SQream Technologies

STMicroelectronics International, N.V.

STREAM

SUSE

Suzhou Supercloud Network Technology Co.

TGM

TECHNOPROM

TIGOCMT

TS8 TECHNOLOGIES

THOMAS KRENN

TVARIT GmbH

Vantosh

Vereign AG

Velankani Electronics Pvt. Ltd.

Wallscope

Xinchangde Science Technology Co., Ltd.

X-ISS eXcellence in IS Solutions

XILINX

XSKY (Beijing) Data Technology Co., Ltd

YiHe Cloud Technology Co.

AntsDB

ZILLIZ

Nallatech

Nanjing Byosoft Co.

Nest Technologies

NEC

Netweb Technologies India PVT Ltd.

NIMBIX

NVIDIA

OCF

Open Computing Singapore PTE LTD

OROBIX TOOLS WITH A MIND

Parabricks

Pengan Cirius

POWERCORE

PLANET Artificial Intelligence GmbH

qBotica

QLOGIC

Quru Limited

rackspace HOSTING

Rambus

RAPTOR ENGINEERING

RECONFIGURE

Red Flag Software Co., Ltd.

REDCOP

RTDS Technologies

S&I Engineering Pvt. Ltd.

Sensor-Me

Shanghai GLOTECH Information Technology Co., Ltd

Shanghai Hongpu Information Technology Co., Ltd.

Shanghai JiSuan Information Technology Company

Kyligence

Infolytics

Ingenious Technologies, LLC

Innovate IT Ltd

Intersil Corporation

IRCONA

Jiangsu ITCAST Educational Technology Co., Ltd.

Kardinal, Inc.

LiveVideo

MAPD

Mellanox

Micron

Microsemi

msi

Mind Research Network

Missing Link Electronics

Mono Cortex

Multicoreware, Inc.

Associate & Academic

A*STAR Computational Resource Centre

AMITY UNIVERSITY DUBAI

J.K.K. Nattraja Educational Institutions

ASU

Aryanet Institute of Technology

ASTRI

AUBURN UNIVERSITY

Barcelona Supercomputing Center

Bauman Moscow State Technical University

BV Raju Institute of Technology (BVRIT)

California Institute of Technology (Caltech)

Carnegie Mellon University

Centre for Advanced Computing

Centre for the Development of Advanced Computing

Clemson University

Clarkson UNIVERSITY

College of Engineering - Anna University

Colorado School of Mines

Computer Network Information Center, Chinese Academy of Sciences

Computing Center FEBRAS

CINECA

cfms

Custom Computing Research Group, Imperial College

Daegu Gyeongbuk Institute of Science & Technology (DGIST)

Dalian University of Technology

Delft University of Technology

Dept. of Comp. Sci. @ Univ. of Brasilia

FreeBSD

Garwood Center for Corporate Innovation

GENCI

GSIC, Titech, Global Scientific Information and Computing Center

Harbin Institute of Technology Weihai

Hartree Centre

Hasso Plattner Institute for Software Engineering

ICM University of Warsaw

Indian Institute of Science (Computational Aerodynamics Lab)

Indian Institute of Science - Dept. of Computer Science & Automation

Sunway University

Texas A&M Engineering Experiment Station

Tsinghua University

UC Berkeley

Universidad de Buenos Aires

UCLM

UNC

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UNICAMP

University of Central Florida

University Corporation for Atmospheric Research (UCAR)

University of Exeter

University of Florida

University of Hawaii

University of Hyderabad

University of Jaffna

University of Miami

University of Michigan

University of Oregon

University of Oxford

University of Peloponnese

University of Singapore

University of Southern California

University of South Carolina

University of Texas (TACC)

Waseda University

Washington University in St. Louis

University of Warwick

ikm

Western Research

Indian Institute of Technology Roorkee

Institute of Communication and Computer Systems (ICCS)

Institute for Dev. & Research in Banking Technology (IDRB)

Institute of Physics

Interaction

International Robotics School of Mytyshi

Istituto Nazionale di Astrofisica (INAF)

JÜLICH

KLE Technological University

KTH Royal Institute of Technology

Kuppam Engineering College

Lawrence Livermore National Laboratory

LSU LOUISIANA STATE UNIVERSITY

rzg RECHEN-ZENTRUM GARCHING

National Center for High-Performance Computing

National University of Singapore

NCI AUSTRALIA

NIT Mangalore

NIT Warangal

Northeastern University in China

OAK RIDGE National Laboratory

Ohio State University

OSU Oregon State UNIVERSITY

POSTECH

PADERBORN UNIVERSITY

PSG College of Technology

PPD Power Progress Community

RICE

Rensselaer Polytechnic Institute (RPI)

Romeo HPC Center

SASTRA University

SETI Institute

Seoul National University

SHEELTRON

Shenzhen University

Sri Saytha Sai Institute of Higher Learning

Stevens Institute of Technology

Associate Individual

Troy Anderson

Vinay Agraharkar

Shawn Anastasio

Bilal Anwar

James Grinalds

Amudhan Balasubramanian

Eugenio Bissacott

Hugh Blemings

Gregory Byrd

Matthais Brobbel

Olga Buchonina

Jeffrey Cassidy

Dr. Chuong-Yu Cho

Jae Min, Cho

Gerry Cocco

Greg L. Coggin

Joseph Conte

Jed Deanne

Andrew DeSomma

Dale Elson

Mustafa Erdogon

Professor Hiroshi Esaki

Patrick Fitzgerald

Alex Fok

Robert Gendron

James Grinalds

Rojit Gupta

James Hall

Jan Henke

Teguh Hofstee

Kevin Holland

Roberto Innocenti

Scott Johnson

YF Juan

Narendra Karmakar

Dan Chyi King

Saurabh Shripad Kshirsagar

Abhishek Kumar

Praveen Kumar B.A.

Mageshkumar Kuppusamy

Dr. Eldad Levy

Chris Lockhart

Justin Lynn

Benjamin Kerensa

Saurabh Shripad Kshirsagar

G.S. Madhusudan

Molefi Alfred Makuebu

Jason Mars

Michael Mayerhofer - Germany

Gordon McKean

Kiran Narenda Mehta

Roy Minor

Rajiv Misra

Steven Munroe

Iyyanki Muralikrishna

Takaaki Nakajima

V Krishna Nandivada - India

Matt Nawrocki

Thomas Norgaard

Thoi Nguyen

Marco Olguin

John Owens

Gael Paul

Maria Letizia Piccaro

Daniel Pocock

Pascal Polverini

Abhinandan Sridhara Rao Prasad

Koya Surendra Prasad

Rohini Puttabasavaiah

Kevin Rakos

Bharov Reddy

Javier Romero

Randall Ross

Jason Rotella

William Rowe

Mijo Safradin - Germany

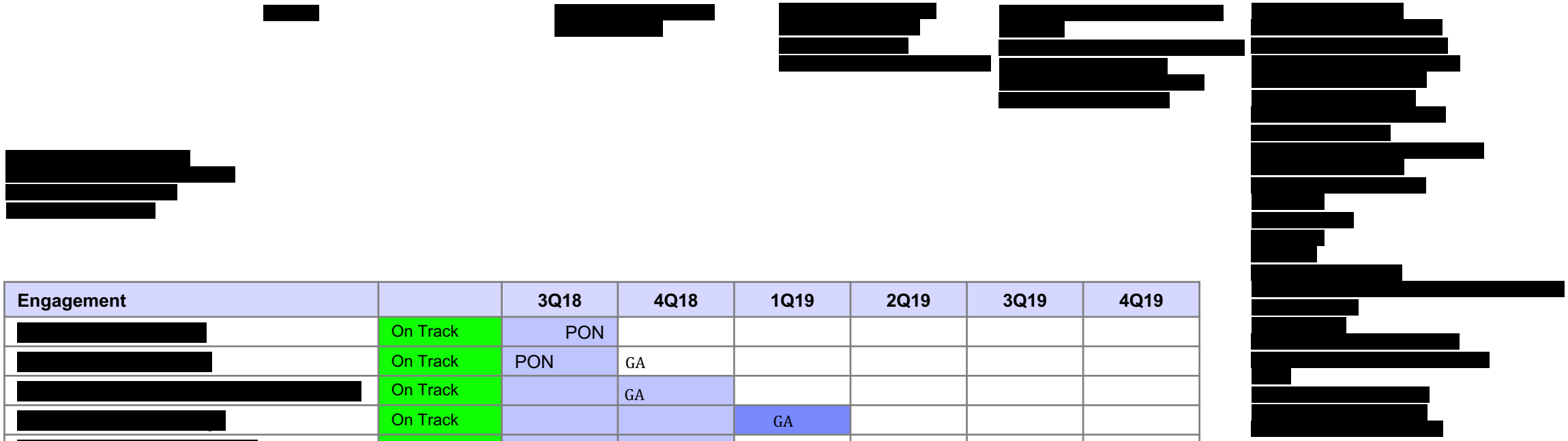
James Sander

Jonas Skeppstedt

https://openpowerfoundation.org

OpenPOWER – Partner Engagement Summary (2018)

New /Pending	Discussion	Concept	Design	Power ON	Ship Ready	Total
1	0	2	3	7	29	39



	Engagement		3Q18	4Q18	1Q19	2Q19	3Q19	4Q19
China	[redacted]	On Track	PON					
	[redacted]	On Track	PON	GA				
	[redacted]	On Track		GA				
	[redacted]	On Track			GA			
	[redacted]	On Track						
	[redacted]	On Track				PON		
ROW	[redacted]	On Track		GA				
	[redacted]	On Track		GA				
	[redacted]	On Track	Concept					
	[redacted]	On Track			GA			
	[redacted]	On Track					PON	

The TWO most POWERFUL HPC systems: Summit & Sierra

The United States Department of Energy together with Oak Ridge National Laboratory and Lawrence Livermore National Laboratory have contracted **IBM and Nvidia** to build two supercomputers, the Summit and the Sierra, that are **based on POWER9 processors coupled with Nvidia's Volta GPUs**. These systems went online in 2018.

IBM Summit #1 !!

IBM Sierra #2 !!

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
2	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
3	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
4	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 , NUDT National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482

IBM – Total Pangea III #5 !!

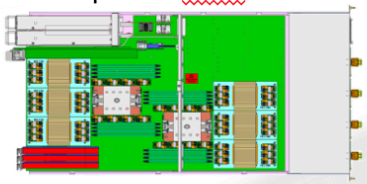
http://www.teratec.eu/actu/calcul/Nvidia_Coral_White_Paper_Final_3_1.pdf

Summit Overview

OpenPOWER™

Compute Node

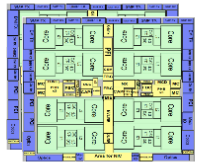
2 x POWER9
6 x NVIDIA GV100
NVMe-compatible PCIe 1600 GB SSD



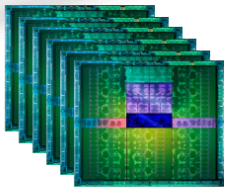
25 GB/s EDR IB- (2 ports)
512 GB DRAM- (DDR4)
96 GB HBM- (3D Stacked)
Coherent Shared Memory

Components

IBM POWER9
- 22 Cores
- 4 Threads/core
- NVLink



NVIDIA GV100
- 7 TF
- 16 GB @ 0.9 TB/s
- NVLink



Compute Rack

18 Compute Servers
Warm water (70°F direct-cooled components)
RDHX for air-cooled components

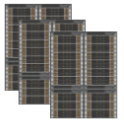
39.7 TB Memory/rack
55 KW max power/rack

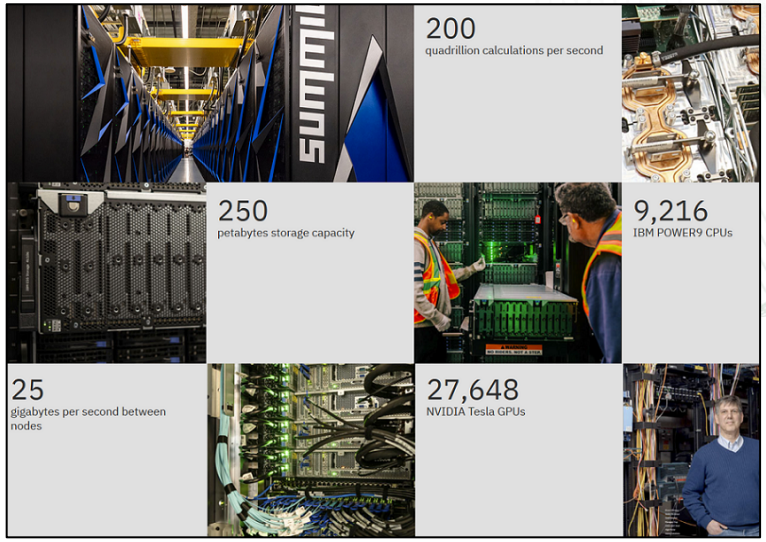
Compute System

10.2 PB Total Memory
256 compute racks
4,608 compute nodes
Mellanox EDR IB fabric
200 PFLOPS
~13 MW

GPFS File System

250 PB storage
2.5 TB/s read, 2.5 TB/s write





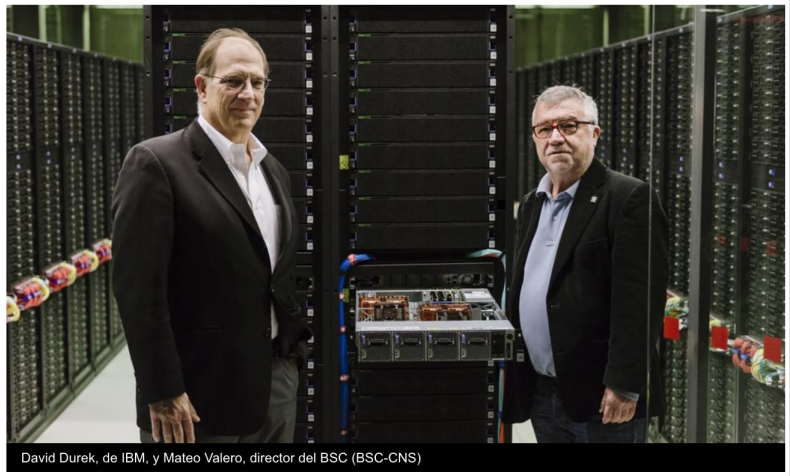
OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

BSC Mare Nostrum 4

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

El MareNostrum 4 amplía su capacidad para investigar en IA

• El BSC se convierte en el primer centro de Europa en ofrecer acceso a las mismas tecnologías que el nuevo supercomputador Summit, de EE.UU., el más potente del mundo



David Durek, de IBM, y Mateo Valero, director del BSC (BSC-CNS)

Part of Mare Nostrum 4

- 3 Racks
- 54 Power9 Systems
  - 54 AC922 servers
  - 4x Nvidia V100
  - 512 GB RAM
- 6.4 TB NVMe storage
- 1.48 Pflops !

IBM's Pangea III is the world's most powerful commercial supercomputer

Total's Supercomputer Ranked First in Industry Worldwide

The new **IBM POWER9-based supercomputer (25 PFLOPS & 50 Pbytes)** will help Total more accurately locate new resources and better assess the potential of new opportunities.

According to Total **Pangea III requires 1.5 Megawatts, compared to 4.5 MW for its predecessor system**. Combined with the increased performance of Pangea III, Total has reported that they have observed that the new system **uses less than 10% the energy consumption per petaflop as its predecessor**.

- **Higher Resolution Seismic Imaging in exploration and development phase**
- **Reliable Development and Production Models**
- **Asset Valuation and Selectivity**

MIT SATORI "Sudden Enlightment"

IBM POWER9 + NVIDIA

https://researchcomputing.mit.edu/satori/home/

satori.mit.edu is the name of a new scalable AI oriented hardware resource for research computing at MIT. It is made possible by a donation through IBM Global Universities Program. Provided as a gift from IBM it will help further the aims of the new MIT Stephen A. Schwarzman College of Computing and other campus initiatives that are combining supercomputing power and AI algorithmic innovation.

- 64 IBM Power 9 Nodes
- 256 NVidia V100 GPUs
- EDR Infiniband
- 2PB storage
- 8TB GPU memory
- 64 TB main memory
- IBM Power AI Software
- Cloud Integration

?

IBM POWER9 + NVIDIA

?

?

IBM® Power System™ Accelerated Compute Server (AC922)

The Best Server for Enterprise AI

POWER9

An acceleration superhighway.
The only processor specifically designed for the AI era.

4x
Threads per core vs x86

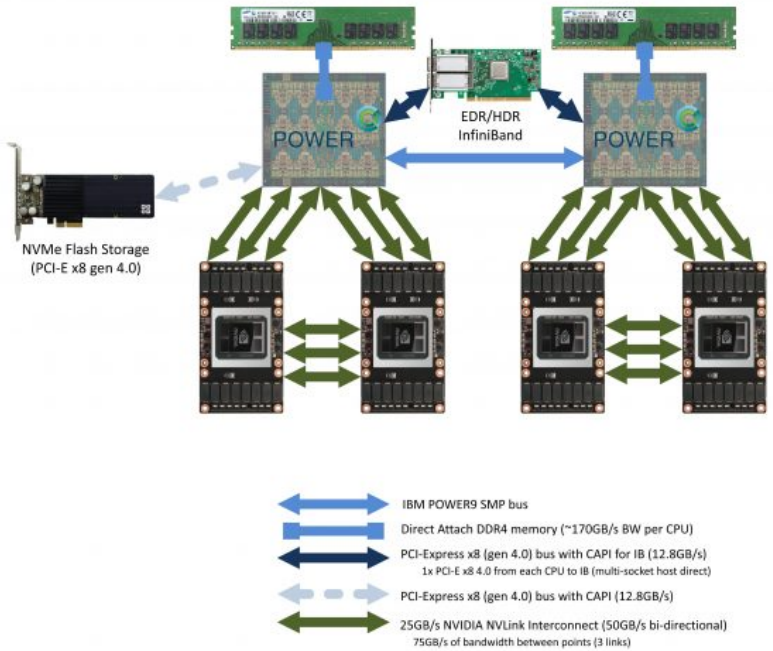
9.5x
Up to 9.5x more I/O bandwidth than x86

2.6x
More RAM possible vs. x86

1st
CPU to deliver PCIe gen 4

Server Block Diagram

Power Systems AC922 with NVIDIA Tesla V100 with Enhanced NVLink GPUs



Innovation from an ecosystem of partners across the stack and open to the core

OpenPOWER™ READY

OPEN POWER

250+ OpenPOWER members co-design around the core to accelerate cognitive and general workloads

NVIDIA, XILINX, Micron, Mellanox TECHNOLOGIES, +++

OPEN SOURCE

Density, speed-up, compaction of the most ubiquitous open source engines

redis, EDB POSTGRES, PostgreSQL, cassandra, MariaDB, neo4j, mongoDB., Linux, Spark, MySQL, +++

OPEN CAPI

Laying the groundwork for faster coherent open interfaces to attach to accelerators

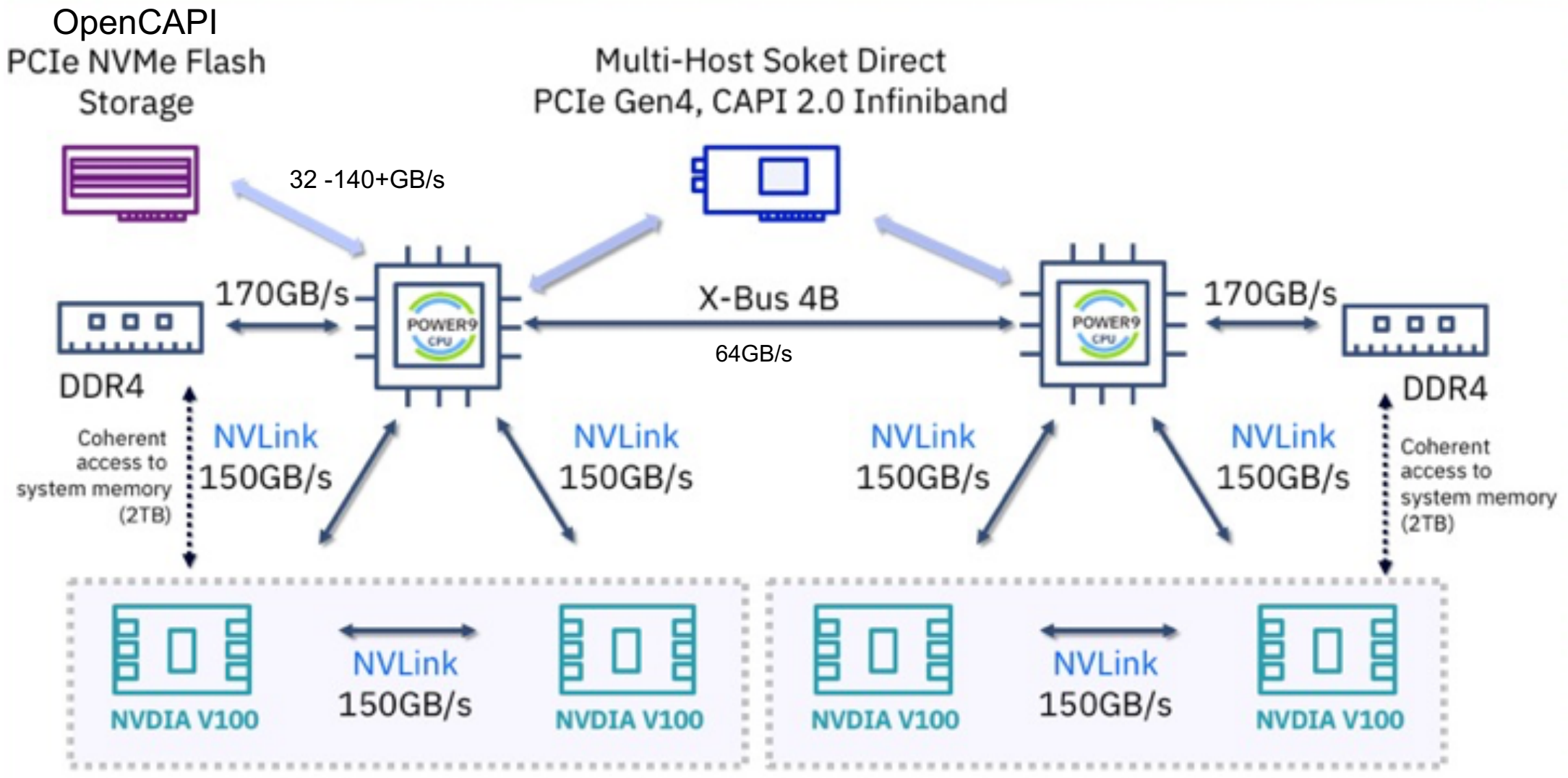
Google, AMD, DELL, Hewlett Packard Enterprise, +++

OPEN FRAMEWORKS

The industry's most ubiquitous Cognitive/AI frameworks - optimized and accelerated.

Caffe, DL4J DEEPLEARNING4J, torch, Chainer, theano, TensorFlow, +++

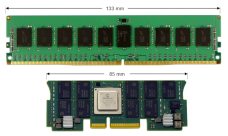
AC922 System buses and components diagram

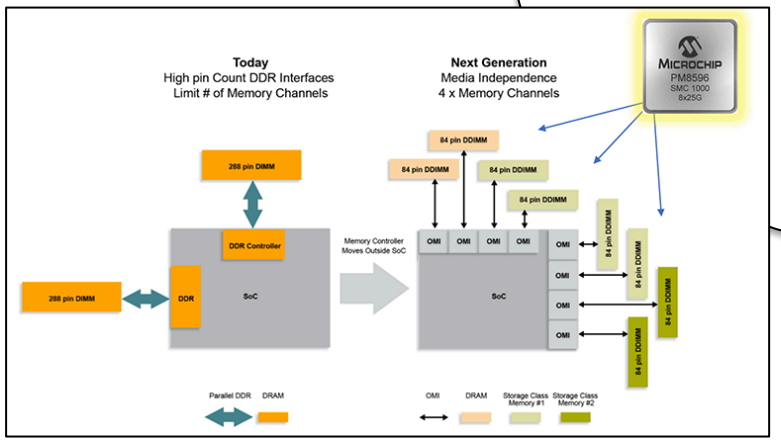


Open POWER moves to Linux Foundation
IBM Demonstrates Commitment to Open Hardware Movement

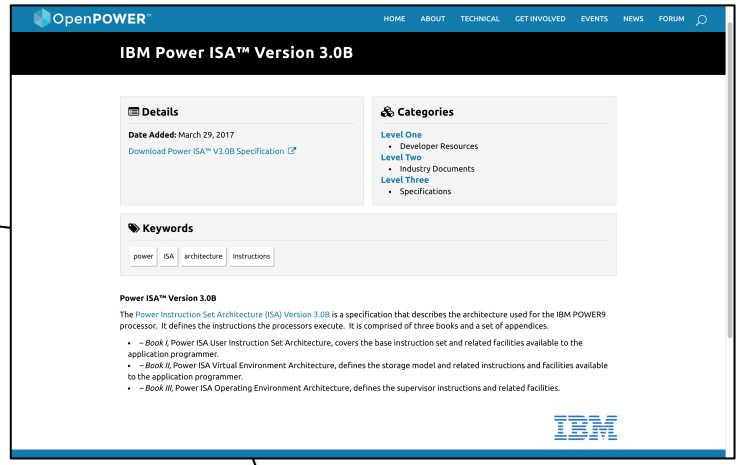
https://openpowerfoundation.org/?resource_lib=power-isa-version-3-0

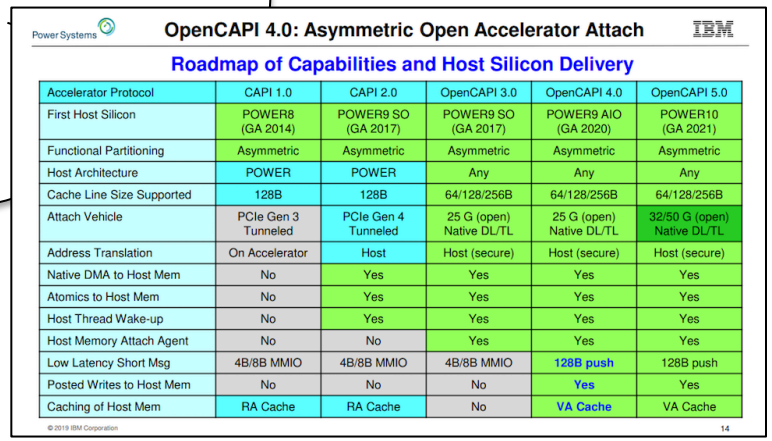
- **Open POWER ISA** (instruction set architecture)
  - **Microwatt** (Power ISA soft core)
- **Open CAPI** (coherent acelerator processor interface)
- **OMI** (Open Memory Interface)





THE LINUX FOUNDATION





Accelerator Protocol	CAPI 1.0	CAPI 2.0	OpenCAPI 3.0	OpenCAPI 4.0	OpenCAPI 5.0
First Host Silicon	POWER8 (GA 2014)	POWER9 SO (GA 2017)	POWER9 SO (GA 2017)	POWER9 AIO (GA 2020)	POWER10 (GA 2021)
Functional Partitioning	Asymmetric	Asymmetric	Asymmetric	Asymmetric	Asymmetric
Host Architecture	POWER	POWER	Any	Any	Any
Cache Line Size Supported	128B	128B	64/128/256B	64/128/256B	64/128/256B
Attach Vehicle	PCIe Gen 3 Tunneled	PCIe Gen 4 Tunneled	25 G (open) Native DL/TL	25 G (open) Native DL/TL	32/50 G (open) Native DL/TL
Address Translation	On Accelerator	Host	Host (secure)	Host (secure)	Host (secure)
Native DMA to Host Mem	No	Yes	Yes	Yes	Yes
Atomics to Host Mem	No	Yes	Yes	Yes	Yes
Host Thread Wake-up	No	Yes	Yes	Yes	Yes
Host Memory Attach Agent	No	No	Yes	Yes	Yes
Low Latency Short Msg	4B/8B MMIO	4B/8B MMIO	4B/8B MMIO	128B push	128B push
Posted Writes to Host Mem	No	No	No	Yes	Yes
Caching of Host Mem	RA Cache	RA Cache	No	VA Cache	VA Cache

https://openpowerfoundation.org/the-next-step-in-the-openpower-foundation-journey/
https://www.talospace.com/2019/09/a-beginners-guide-to-hacking-microwatt.html

OpenCAPI

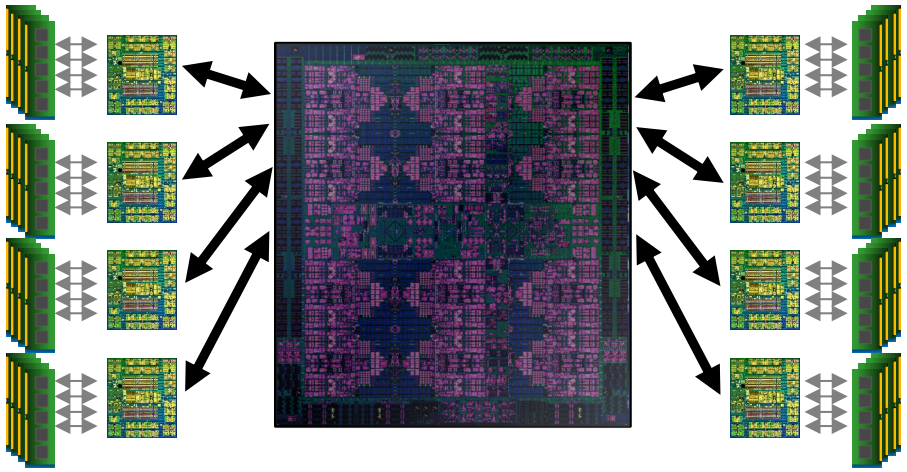
- **What is OpenCAPI?**
  - OpenCAPI is an Open Interface Architecture that allows any microprocessor to attach to
    - Coherent user-level accelerators and I/O devices
    - Advanced memories accessible via read/write or user-level DMA semantics
    - Agnostic to processor architecture
- **Key Attributes of OpenCAPI**
  - High-bandwidth, low latency interface optimized to enable streamlined implementation of attached devices
    - 25Gbit/sec signaling and protocol built to enable very low latency interface on CPU and attached device
    - Complexities of coherence and virtual addressing implemented on host microprocessor to simplify attached devices and facilitate interoperability across multiple CPU architectures
  - Attached devices operate natively within an application's user space and coherently with processors
    - Allows attached device to fully participate in application without kernel involvement/overhead
  - Supports a wide range of use cases and access semantics
    - Hardware accelerators
    - High-performance I/O devices
    - Advanced memories
  - 100% Open Consortium / All company participants welcome / All ISA participants welcome

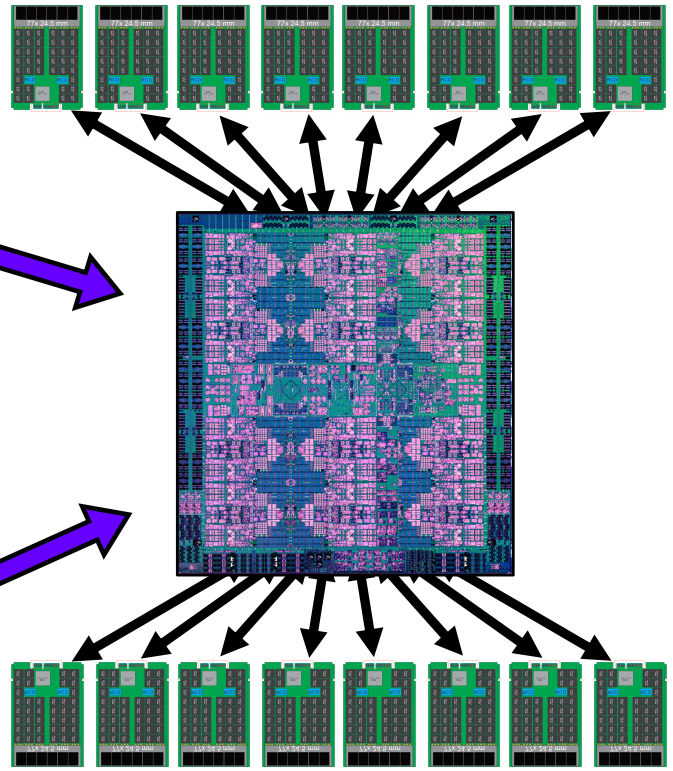
POWER9 Family Memory Architecture

Scale Up
Buffered Memory

Superior RAS, High bandwidth, High Capacity

Agnostic interface for alternate memory innovations

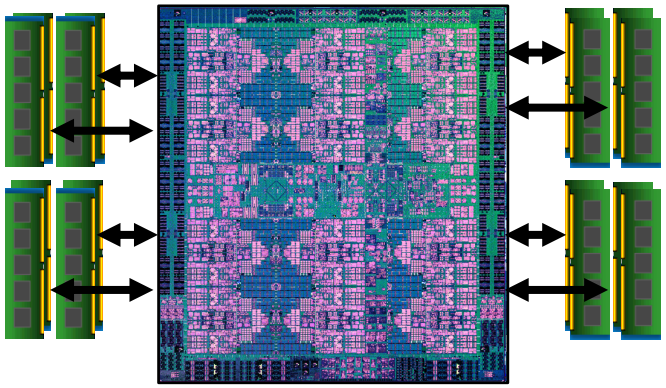




Scale Out
Direct Attach Memory

Low latency access

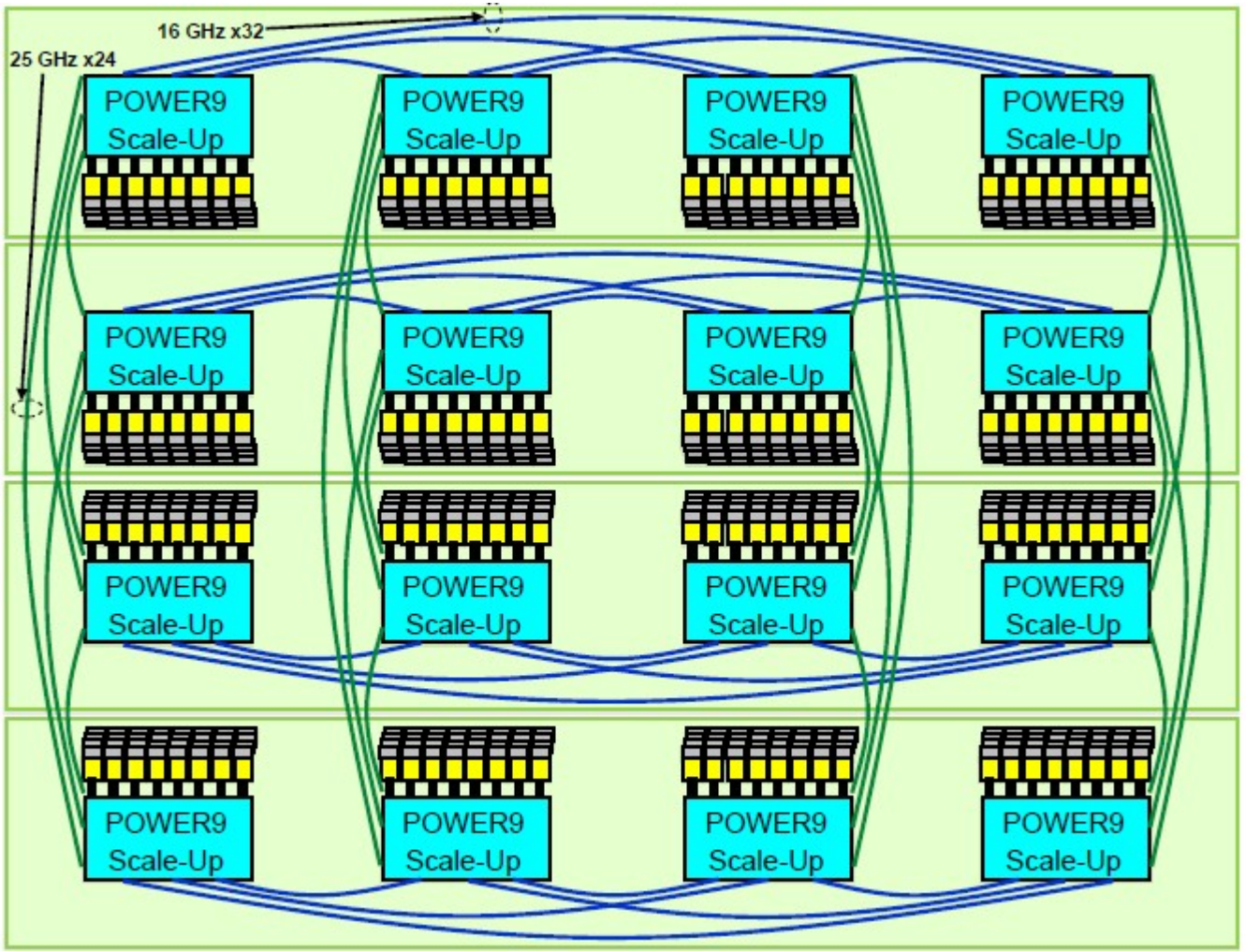
Commodity packaging form factor



OpenCAPI Agnostic Buffered Memory "OMI"

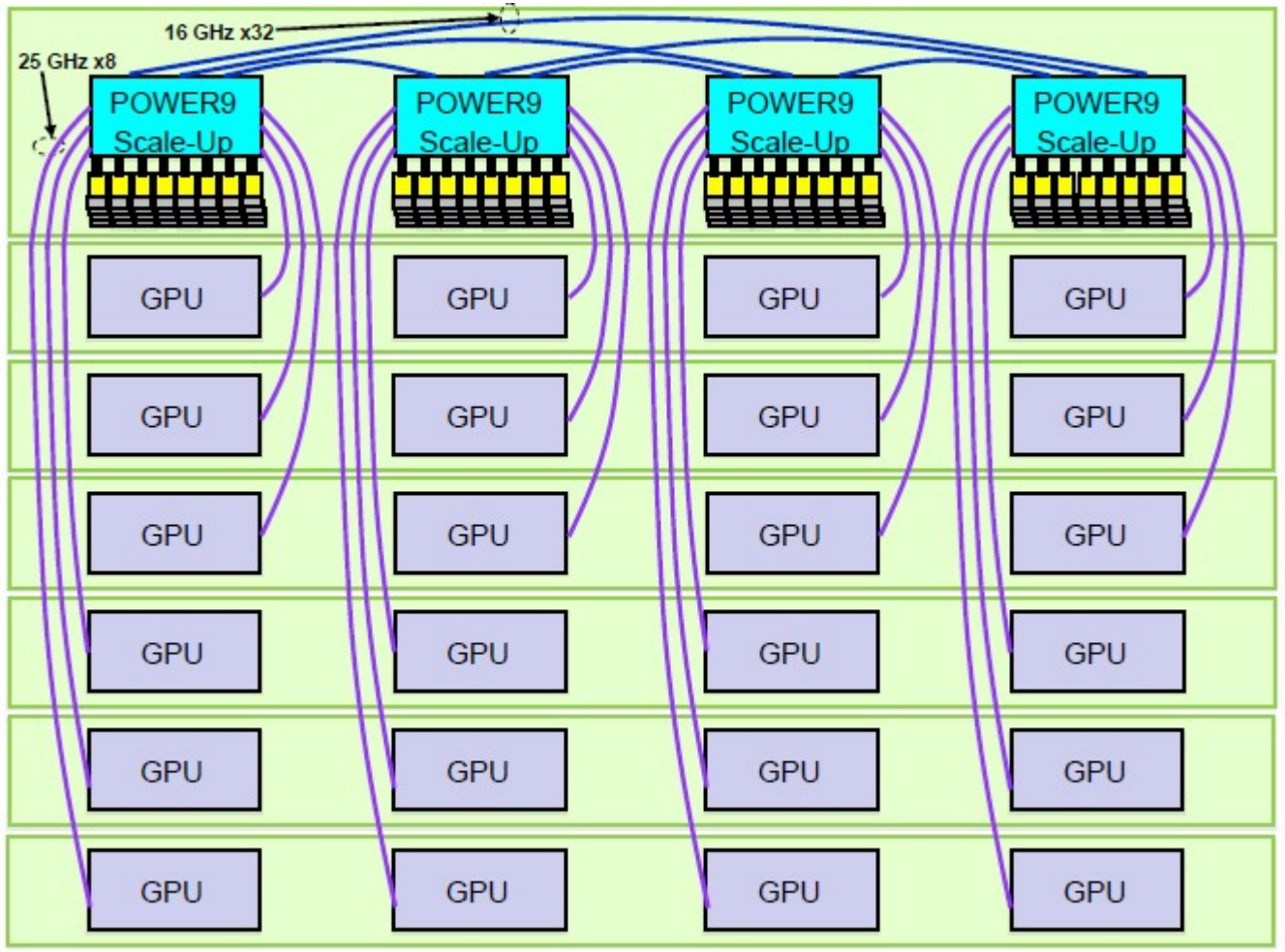
Near Tier	Commodity	Enterprise	Storage Class
Extreme Bandwidth	Low Latency	RAS	Extreme Capacity
Low Capacity	Low Cost	Capacity	Persistence
		Bandwidth	

Somewhere in a possible future… (16 ways POWER9 server using OpenCAPI?)



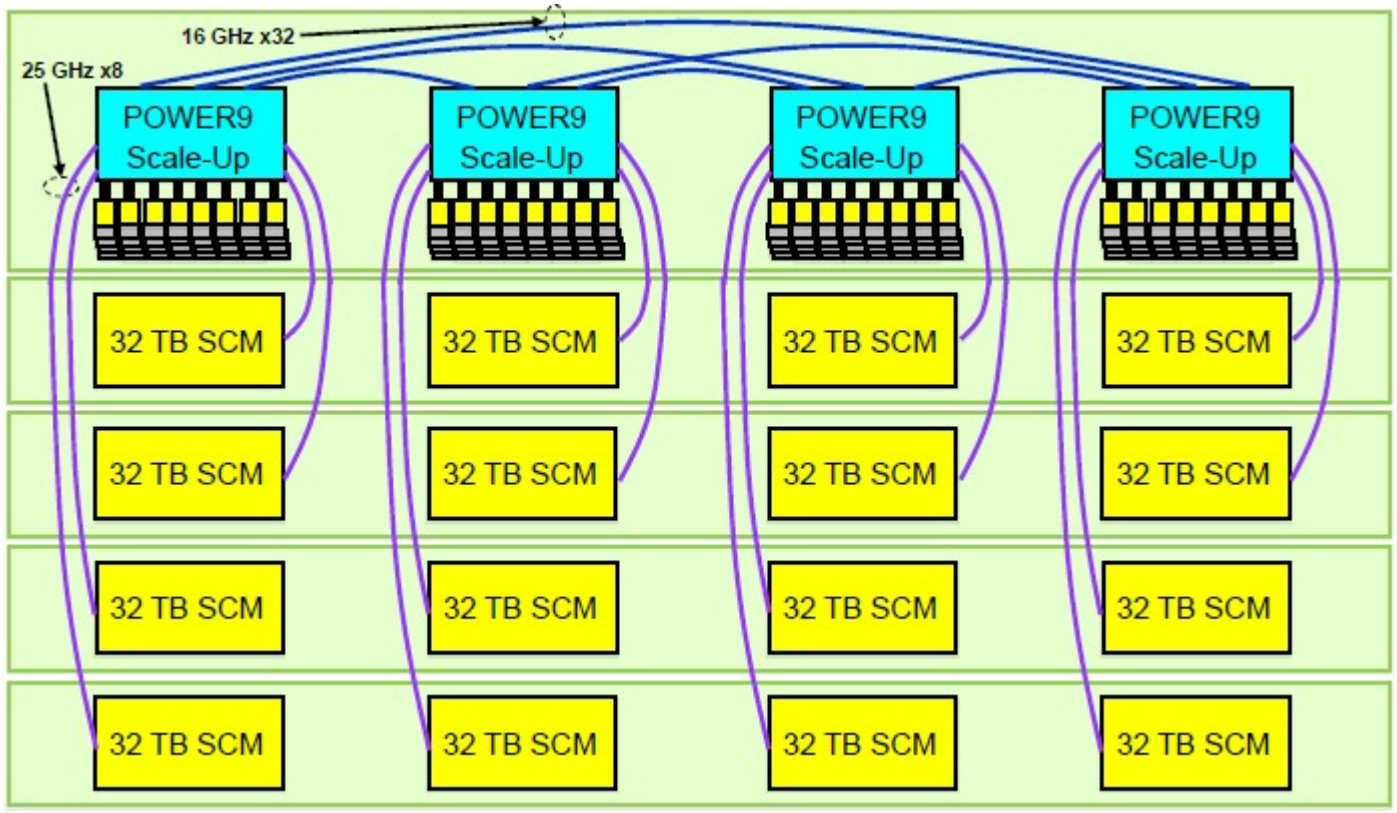
https://www.nextplatform.com/2018/08/28/ibm-power-chips-blur-the-lines-to-memory-and-accelerators/

Somewhere in a possible future… (24 GPU server using OpenCAPI/NVLINK?)



https://www.nextplatform.com/2018/08/28/ibm-power-chips-blur-the-lines-to-memory-and-accelerators/

Somewhere in a possible future… (512TB Storage Class Memory using OpenCAPI?)



https://www.nextplatform.com/2018/08/28/ibm-power-chips-blur-the-lines-to-memory-and-accelerators/

Somewhere in a possible future…

OpenPOWER X

"The Only Stupid Question is the One You Don't Ask"