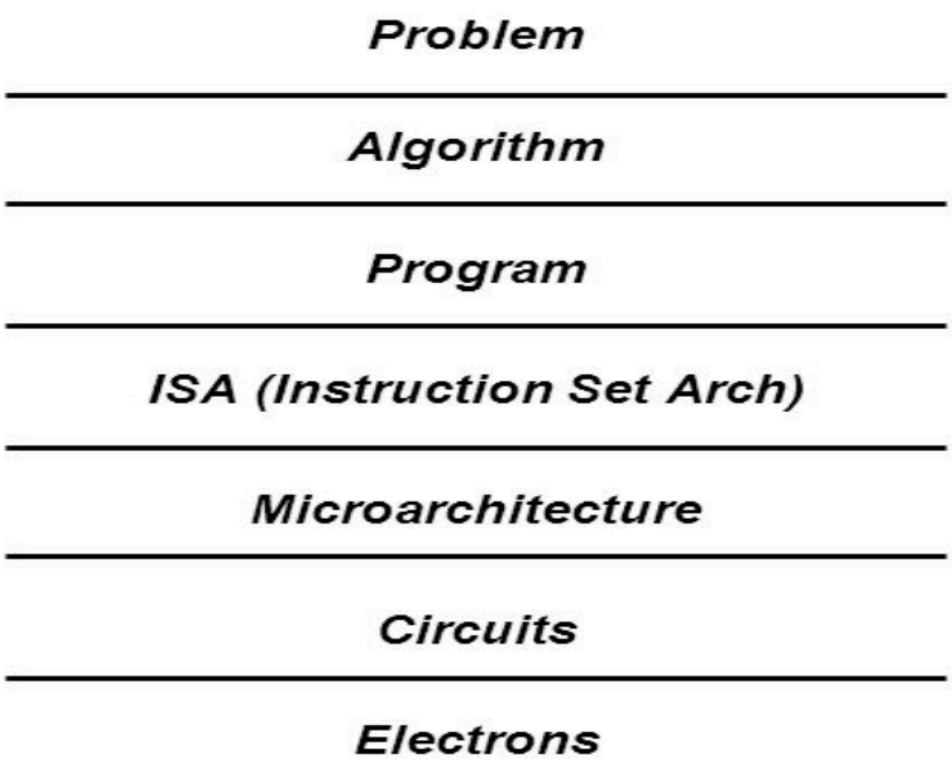
Shared-Memory Heterogeneous Computing

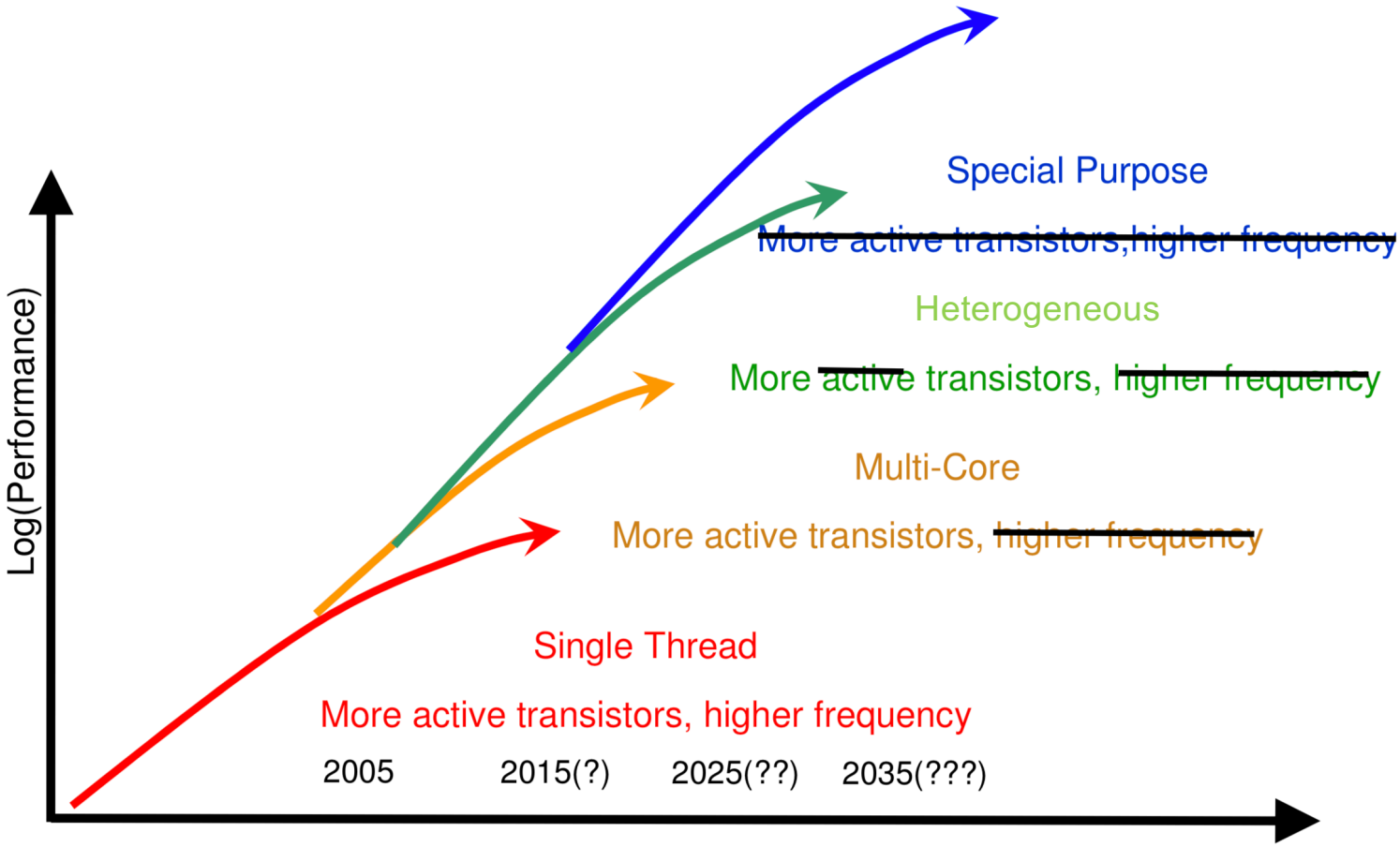
H. Peter Hofstee, Ph.D.
IBM (& TU Delft)

Yale Patt's class April 27, 2020



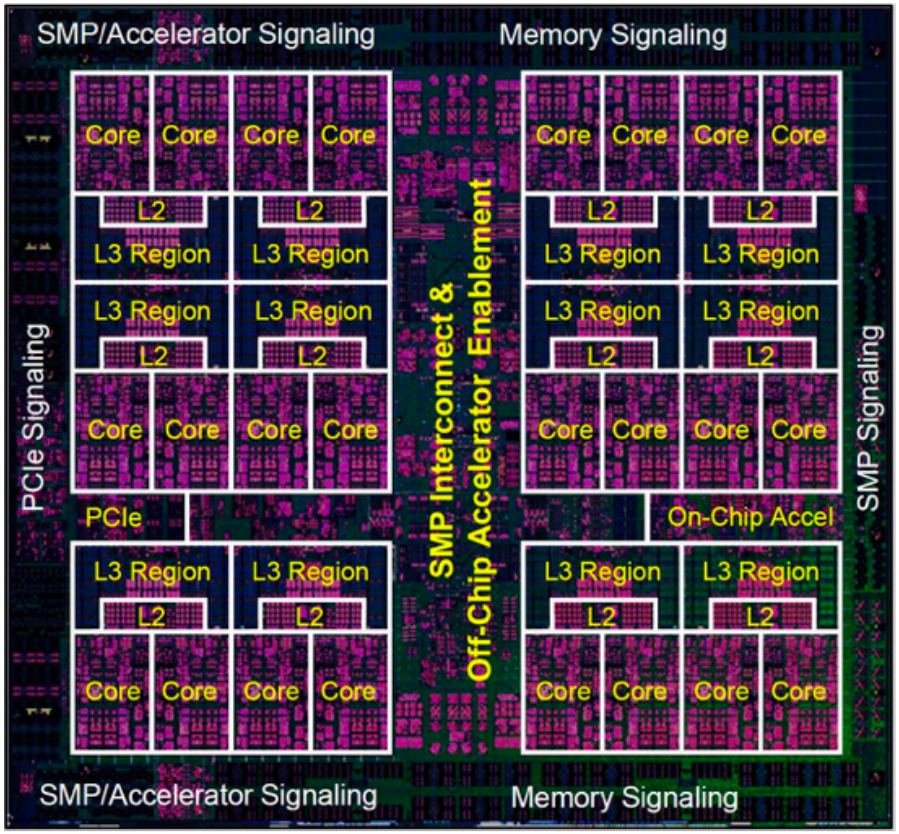
Transformation Hierarchy. Yale Patt

Technology-Driven Processor Trends



after Hofstee
ICCD 2002

Modern CPU

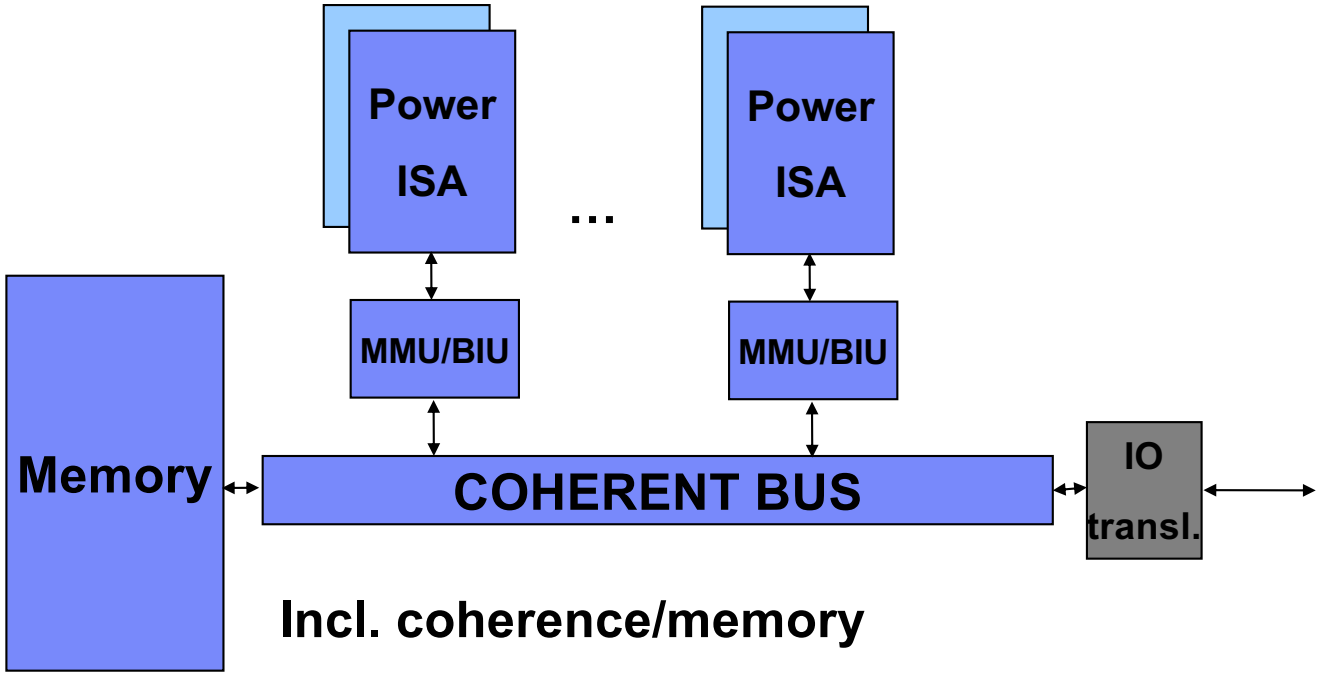


POWER9

Source: "POWER9 Processor for the Cognitive Era". IBM presentation by Brian Thompto. Hot Chips 28 Symposium, October 2016

Cell Architecture is …

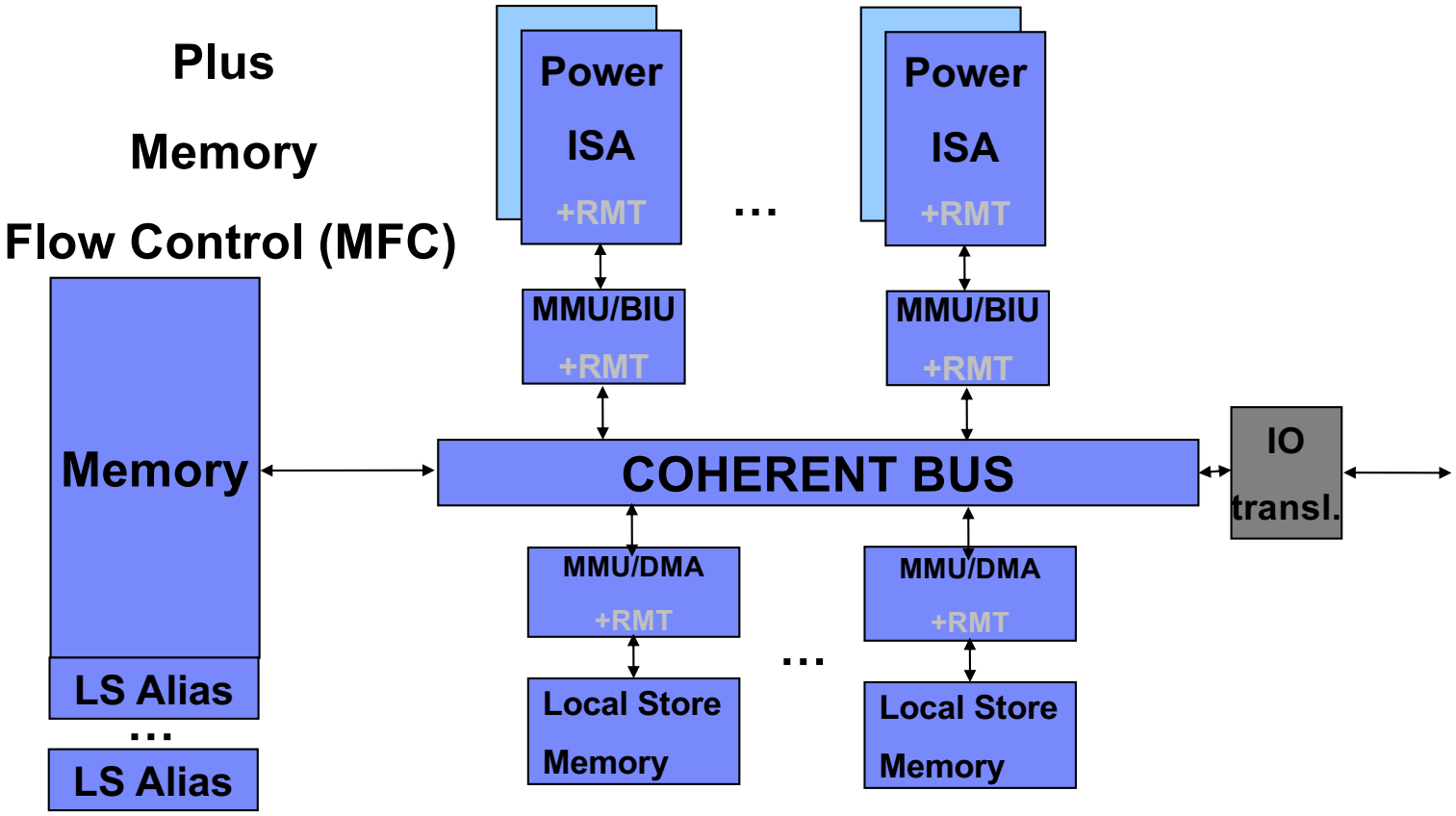
64b Power Architecture™



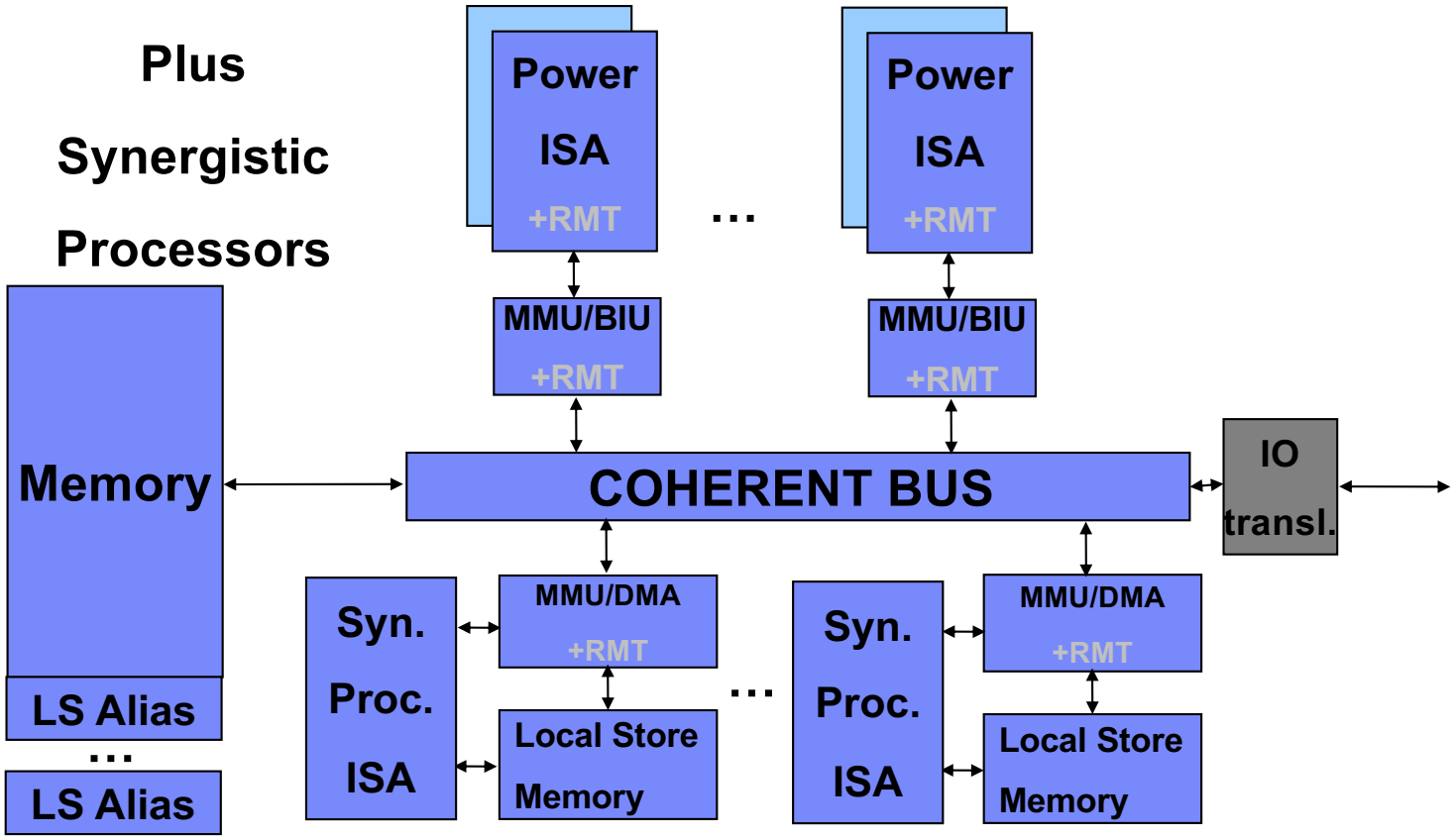
Incl. coherence/memory

compatible with 32/64b Power Arch. Applications and OS's

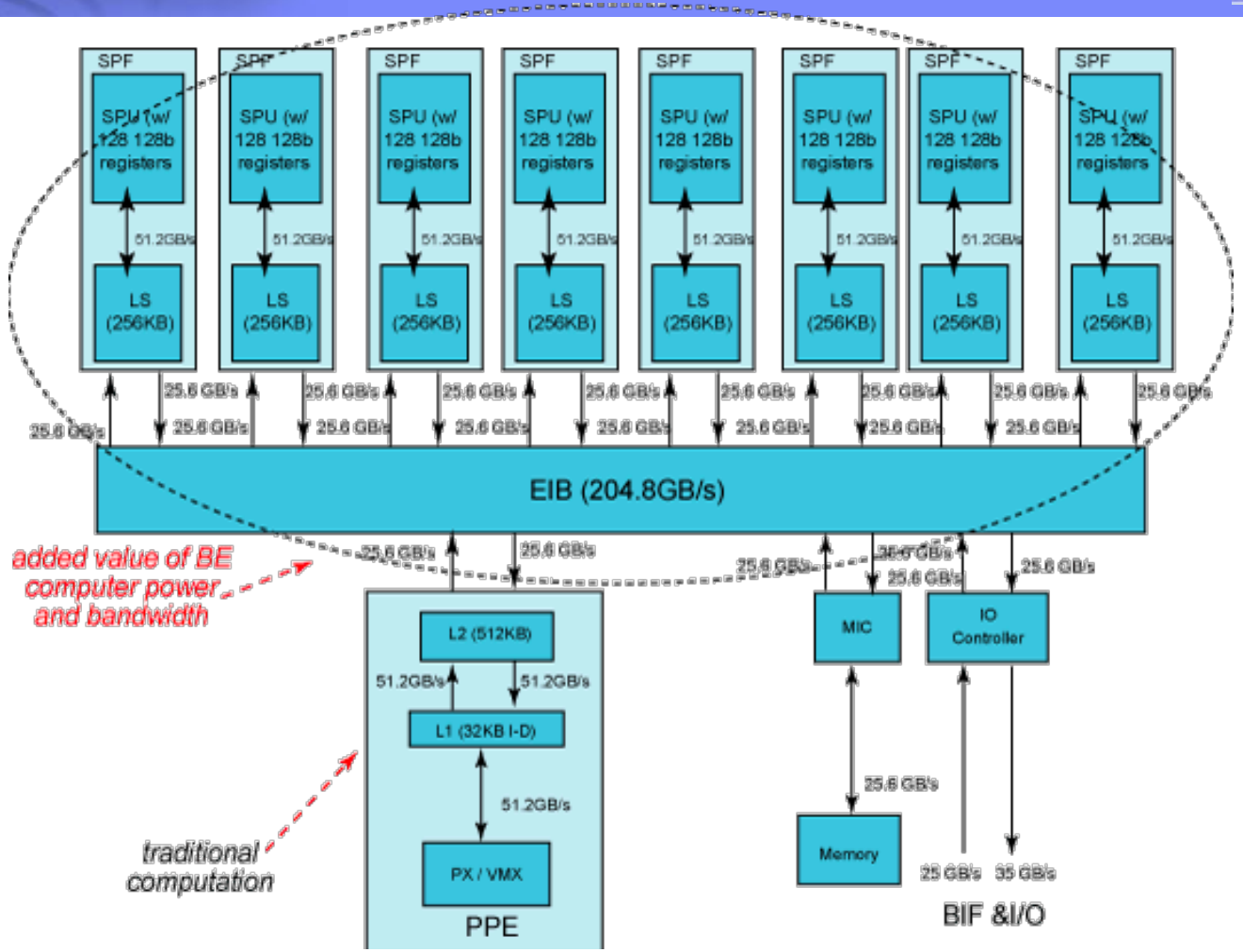
Cell Architecture is … 64b Power Architecture™



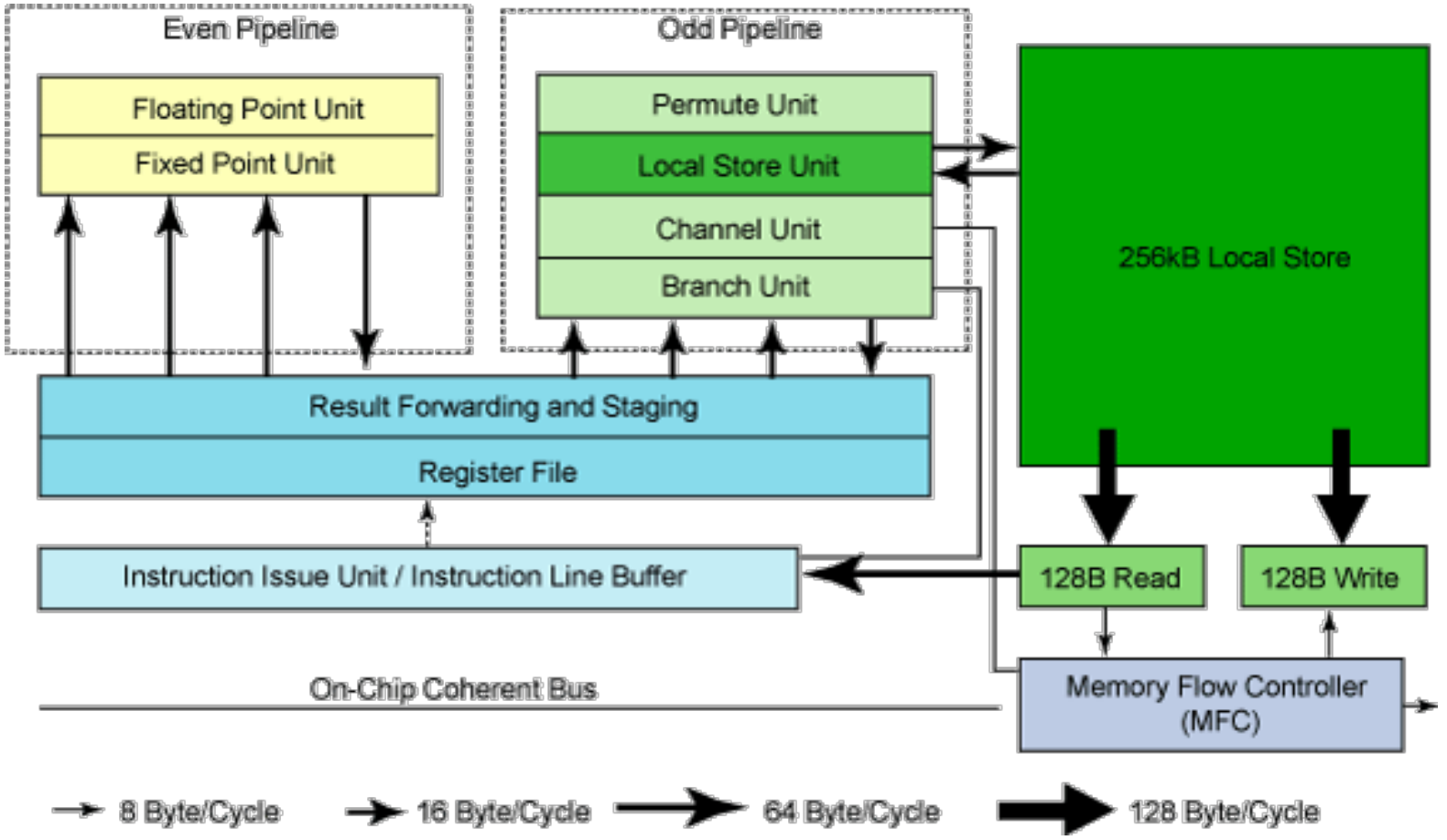
Cell Architecture is … 64b Power Architecture™+ MFC



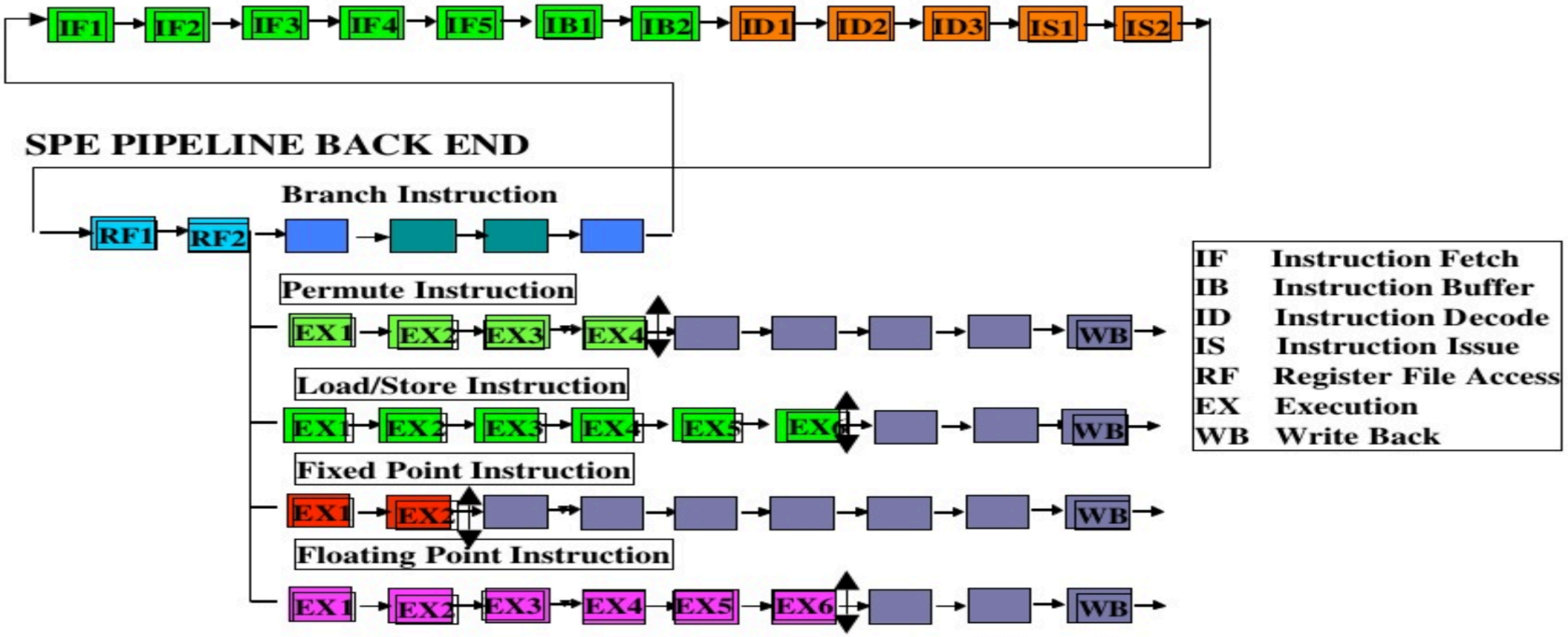
Cell BE specifics



Synergistic Processor Element (SPE) Organization

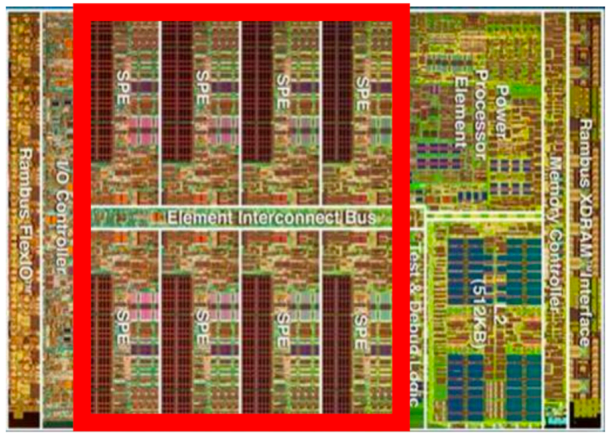


SPU PIPELINE FRONT END

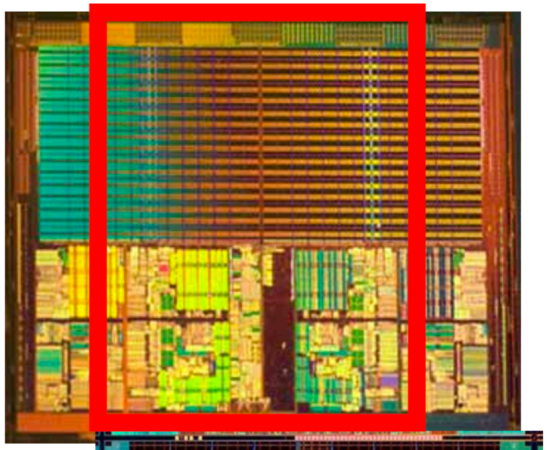


Cell Processor vs. Traditional General Purpose Processor

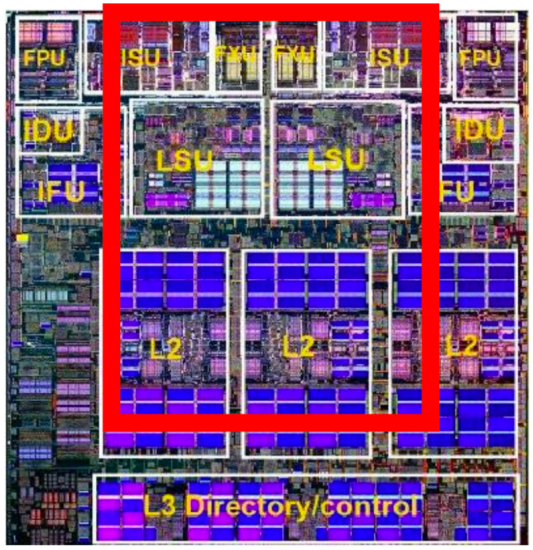
Cell BE



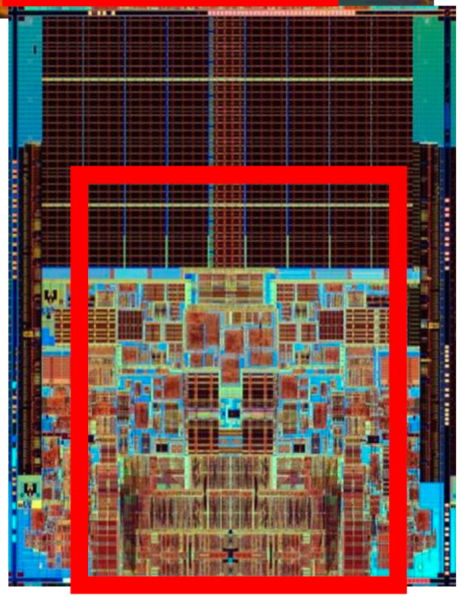
AMD



IBM



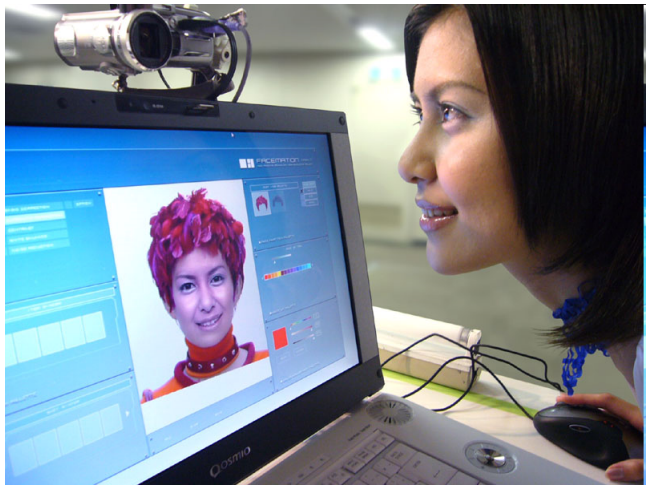
Intel

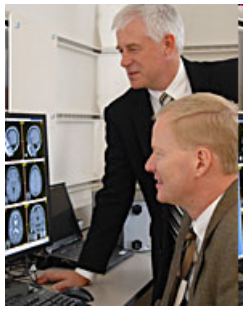


Cell/SPE Architecture Attributes

- Global coherent shared memory
- Local store
- Unified register file
- Asynchronous load/store to local store (dma-list-put/get)
  - Implementation detail: 16 (128B) outstanding shared memory loads & stores per SPE …
- Details … branch hint, wait, …
- Tradeoffs
  - Finite local store size
  - State heavy
  - Local store alias

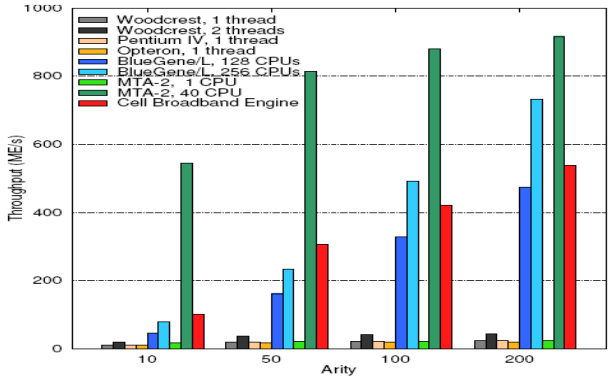
Image/Signal Processing on the Cell Broadband Engine





More Generic Workloads on Cell

Breadth-First Search
Villa, Scarpazza, Petrini, Peinador
IPDPS 2007



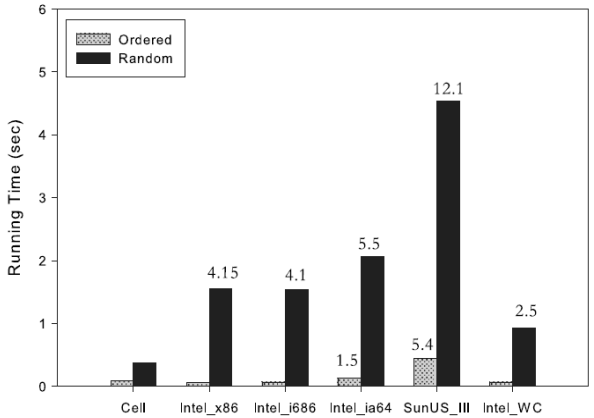
Mapreduce

Sangkaralingam, De Kruijf, Oct. 2007

Application Name	Application Type	Lines of Code		Speedup vs. Core2			BIPS		
		MapReduce	Serial	1-SPE	8-SPEs	8-SPE Ideal	1-SPE	8-SPEs	8-SPE Ideal
histogram	partition-dominated	345	216	0.16	0.15	2.44	1.56	1.51	24.49
kmeans	partition-dominated	324	318	0.91	3.00	6.92	2.08	7.35	17.01
linearRegression	map-dominated	279	114	0.34	2.59	2.67	1.47	11.32	11.70
wordCount	partition-dominated	226	324	0.87	0.96	10.26	1.52	1.74	18.64
NAS_EP	map-dominated	264	112	1.08	8.62	8.62	2.00	15.93	15.95
distributedSort	sort-dominated	171	93[c]	0.41	0.76	5.48	1.28	2.38	17.15

D.A. Bader et al. / Parallel Computing 33 (2007) 720–740

a Comparison of List ranking on Cell with other Single Processors for list of size 8 million nodes



Sort:Gedik, Bordawekar, Yu (IBM)

Table 3: Out-of-core sort performance (in secs)

# items	16 SPEs bitonic	3.2GHz Xeon quick	3.2GHz Xeon quick 2-core	PPE quick
1M	0.0098	0.1813	0.098589	0.4333
2M	0.0234	0.3794	0.205728	0.9072
4M	0.0569	0.7941	0.429499	1.9574
8M	0.1372	1.6704	0.895168	4.0746
16M	0.3172	3.4673	1.863354	8.4577
32M	0.7461	7.1751	3.863495	18.3882
64M	1.7703	14.8731	7.946356	38.7473
128M	4.0991	30.0481	16.165578	79.9971

NOTES ON PROGRAMMING CELL

- Doing it “by hand”
  - DMA-list-get/put
  - Importance of shared memory (L2 in SPE)
  - Synchronization (wake up on lost reservation)

- OpenMP Compilers
  - I-side
  - D-side

- Cray vector machine

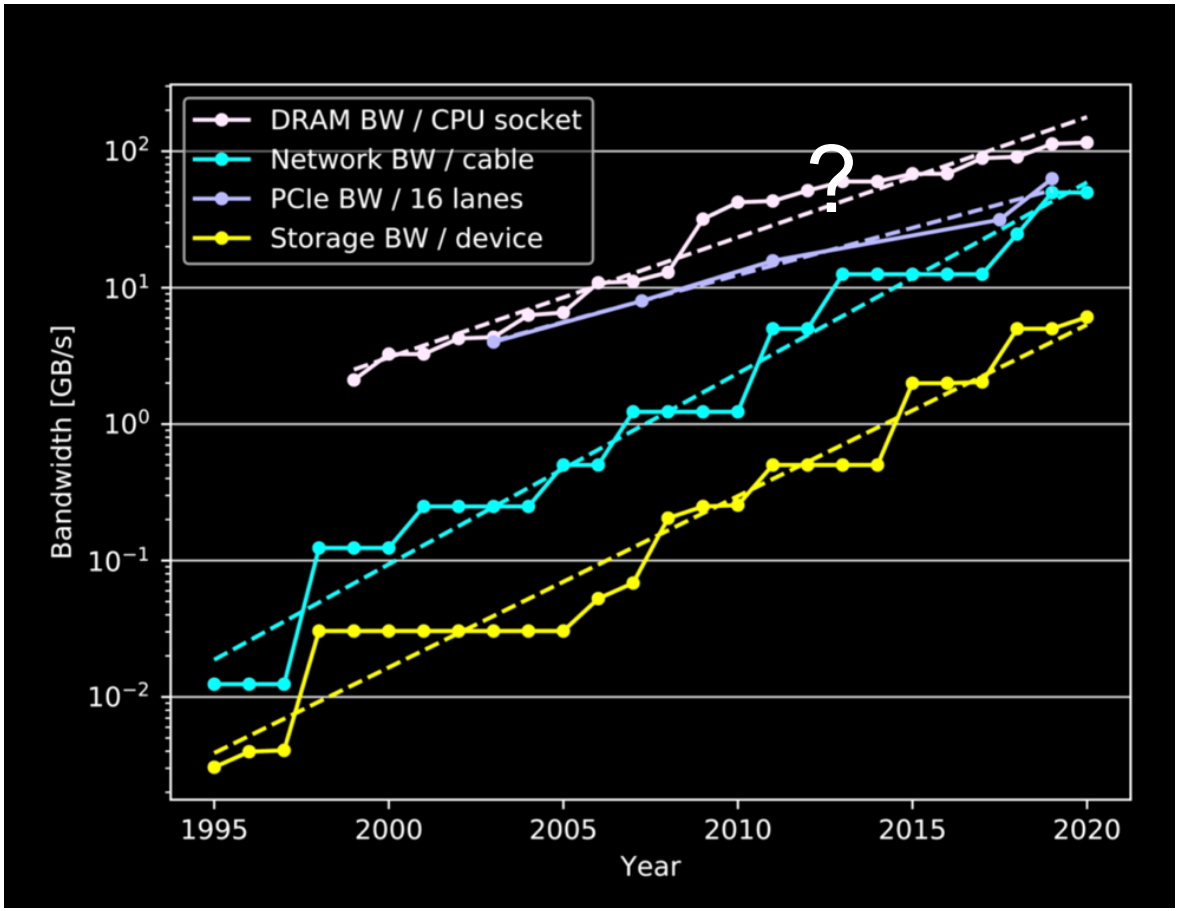
- Tasks on shared memory
  - Many different incarnations
  - Of growing importance in newer versions of OpenMP
  - Local store to local store transfers

- OpenCL and CUDA

CELL Security

- Open architecture
- Some implementation details not disclosed (e.g. key storage and testing)
- Formal verification
- Fundamental decision not to trust the hypervisor
- SPE Isolate-load
- Bootstrapping and how it was hacked

System (Bandwidth) Trends



Based on Sandisk Blog

Sortbenchmark.org

Top Results

	Daytona	Indy
Gray	2016, 44.8 TB/min **Tencent Sort** 100 TB in 134 Seconds 512 nodes x (2 OpenPOWER 10-core POWER8 2.926 GHz, 512 GB memory, 4x Huawei ES3600P V3 1.2TB NVMe SSD, 100Gb Mellanox ConnectX4-EN) Jie Jiang, Lixiong Zheng, Junfeng Pu, Xiong Cheng, Chongqing Zhao Tencent Corporation Mark R. Nutter, Jeremy D. Schaub	2016, 60.7 TB/min **Tencent Sort** 100 TB in 98.8 Seconds 512 nodes x (2 OpenPOWER 10-core POWER8 2.926 GHz, 512 GB memory, 4x Huawei ES3600P V3 1.2TB NVMe SSD, 100Gb Mellanox ConnectX4-EN) Jie Jiang, Lixiong Zheng, Junfeng Pu, Xiong Cheng, Chongqing Zhao Tencent Corporation Mark R. Nutter, Jeremy D. Schaub
Cloud	2016, $1.44 / TB **NADSort** 100 TB for $144 394 Alibaba Cloud ECS ecs.n1.large nodes x (Haswell E5-2680 v3, 8 GB memory, 40GB Ultra Cloud Disk, 4x 135GB SSD Cloud Disk) Qian Wang, Rong Gu, Yihua Huang Nanjing University Reynold Xin Databricks Inc. Wei Wu, Jun Song, Junluan Xia Alibaba Group Inc.	2016, $1.44 / TB **NADSort** 100 TB for $144 394 Alibaba Cloud ECS ecs.n1.large nodes x (Haswell E5-2680 v3, 8 GB memory, 40GB Ultra Cloud Disk, 4x 135GB SSD Cloud Disk) Qian Wang, Rong Gu, Yihua Huang Nanjing University Reynold Xin Databricks Inc. Wei Wu, Jun Song, Junluan Xia Alibaba Group Inc.
Minute	2016, 37 TB **Tencent Sort** 512 nodes x (2 OpenPOWER 10-core POWER8 2.926 GHz, 512 GB memory, 4x Huawei ES3600P V3 1.2TB NVMe SSD, 100Gb Mellanox ConnectX4-EN) Jie Jiang, Lixiong Zheng, Junfeng Pu, Xiong Cheng, Chongqing Zhao Tencent Corporation Mark R. Nutter, Jeremy D. Schaub	2016, 55 TB **Tencent Sort** 512 nodes x (2 OpenPOWER 10-core POWER8 2.926 GHz, 512 GB memory, 4x Huawei ES3600P V3 1.2TB NVMe SSD, 100Gb Mellanox ConnectX4-EN) Jie Jiang, Lixiong Zheng, Junfeng Pu, Xiong Cheng, Chongqing Zhao Tencent Corporation Mark R. Nutter, Jeremy D. Schaub
Joule 10^{10} recs	2-way tie: 2019, 163 KJoules **TaichiSort** 61 K records sorted / joule Intel i7-9700, 32GB RAM, Nsort, Ubuntu 16.04.3 LTS, 2 Intel DC 3600 series PCIe NVMe SSD (1.2 TB), 1 Intel DC 3600 series PCIe NVMe SSD (2.0 TB) Ming Liu, Kaiyuan Zhang, Arvind Krishnamurthy University of Washington Simon Peter University of Texas at Austin 2013, 168 KJoules **NTOSort** 59 K records sorted / joule Intel i7-3770K, 16GB RAM, Nsort, Windows 8, 16 Samsung 840 Pro 256GB SSDs, 1 Samsung 840 Pro 128GB SSD Andreas Ebert Microsoft	2019, 89 KJoules **KioxiaSort** 112 K records sorted / joule Intel i9-9900K, 64GB RAM, Ubuntu 19.04 Server, 8 CFD CSSD-M2B1TPG3VNF (1TB), 1 Toshiba XG5-P KXG50PNV2T04 (2TB) Shintaro Sano, Tomoya Suzuki Kioxia Corporation Zaid Mahmoud Princess Sumaya University for Technology

Notes:

Conventional wisdom: **I/O** limited

Conventional server
- 2POWER8 CPU
- 16x DDR3 DIMM
- 4x NVMe (Gen3 x4)
- 100Gb Ethernet

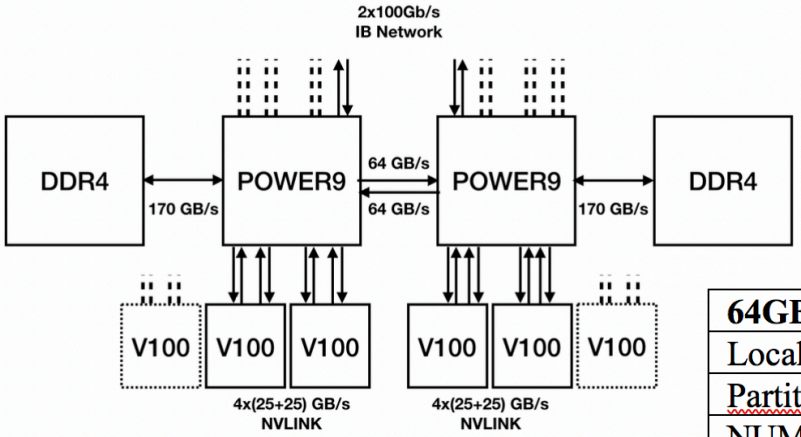
100TB on ~500 servers in 100sec
- ~2GB/s server

Could easily build server @10x I/O
- **COMPUTE limited!**

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]

ı, Texas, USA
ıghai, China
etherlands
ɔm.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

Another Example: Big Data Queries

A lot of information is captured for a lot of your interactions online

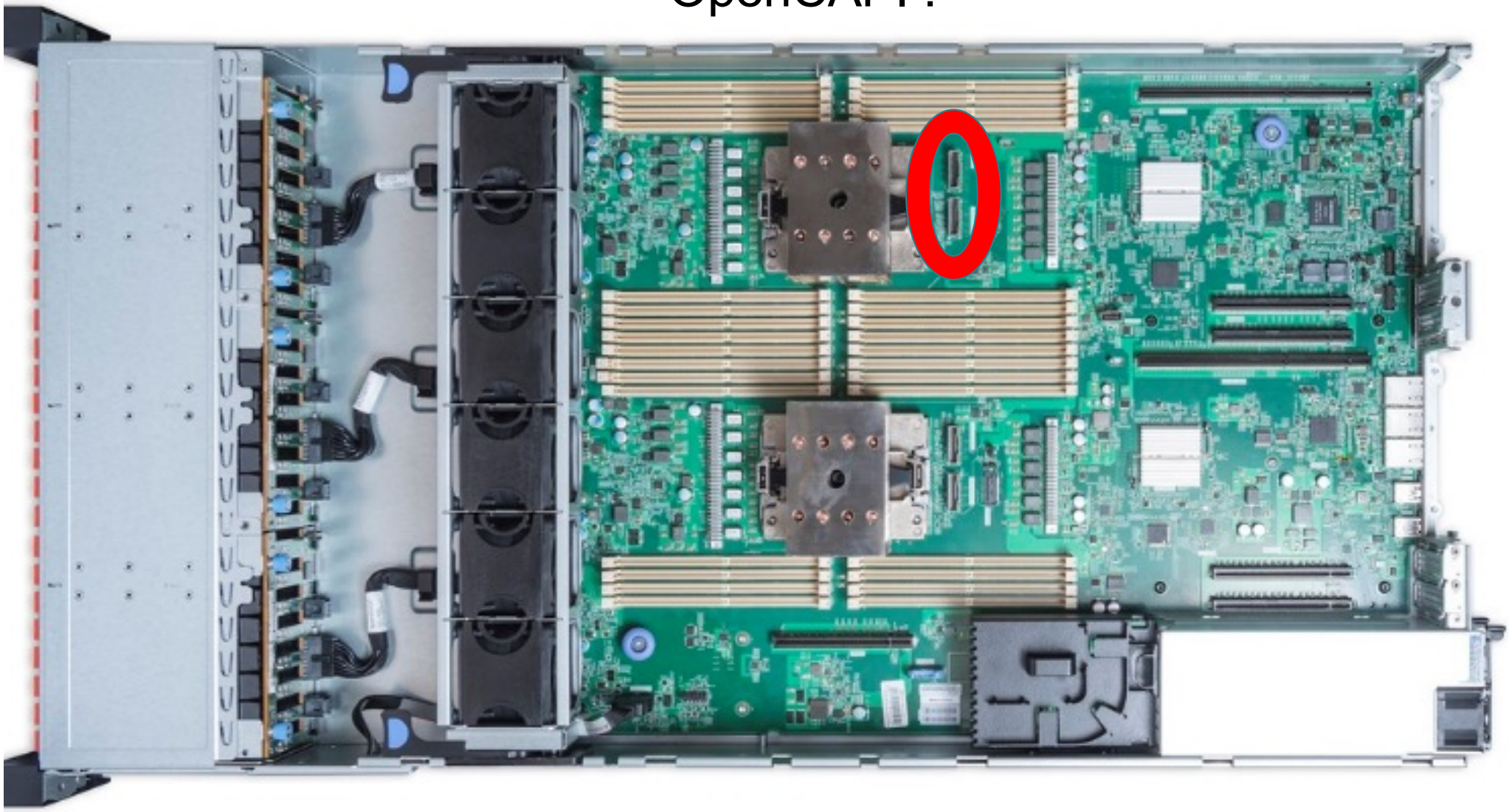
Typically stored in per-column (double) compressed files (ORC/Parquet/ …)

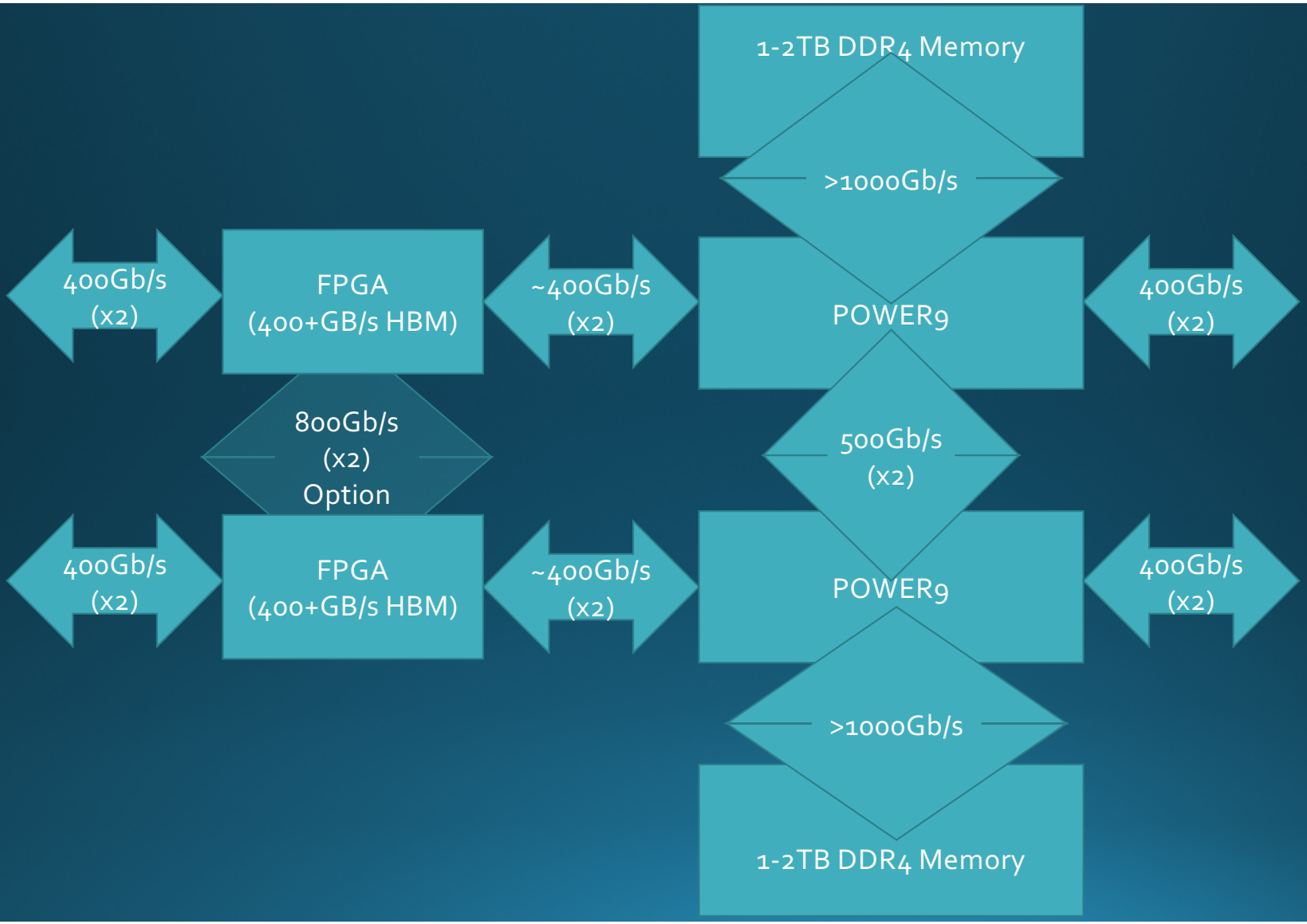
Typical server today can process only at a few GB/s, even for simple queries

Example problems

- Decompression – Would like to do 100GB/s but typical rate is few 100MB/s/core
- Deserialization – In-memory formats differ from on-storage formats, conversion can be costly

OpenCAPI !



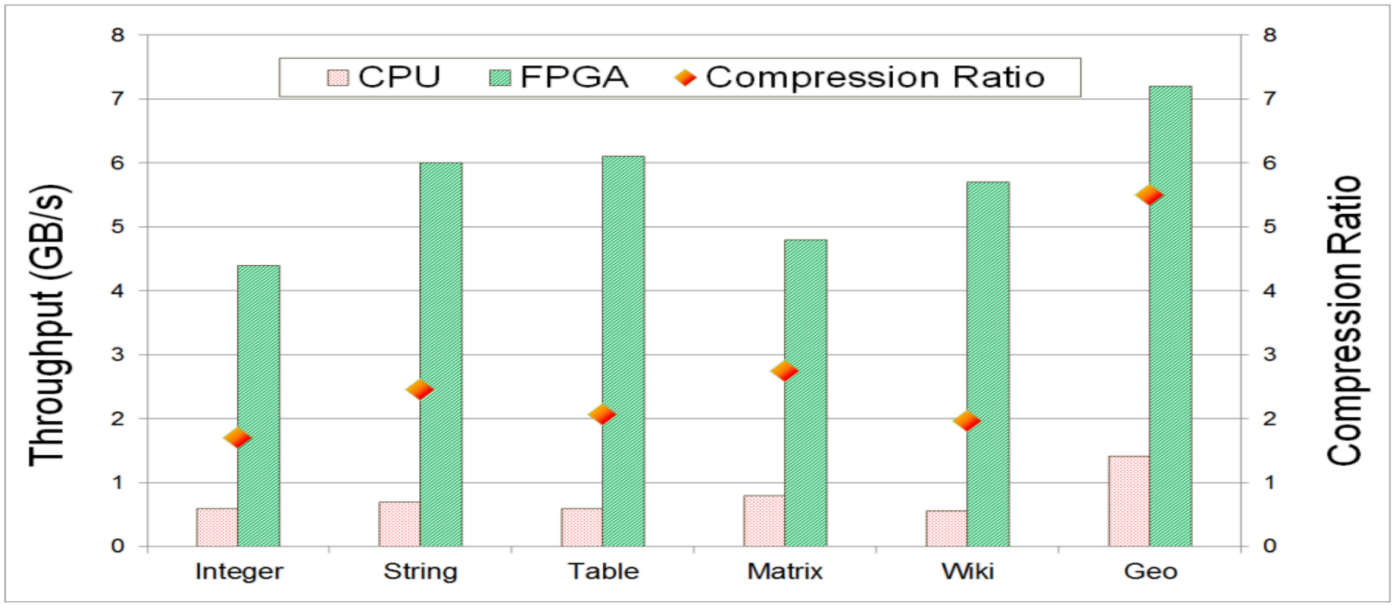

1-2TB DDR4 Memory
>1000Gb/s
400Gb/s
(x2)
FPGA
(400+GB/s HBM)
~400Gb/s
(x2)
POWER9
400Gb/s
(x2)
800Gb/s
(x2)
Option
500Gb/s
(x2)
400Gb/s
(x2)
FPGA
(400+GB/s HBM)
~400Gb/s
(x2)
POWER9
400Gb/s
(x2)
>1000Gb/s
1-2TB DDR4 Memory

Experiment - Results

- End-to-end Throughput

Resource	LUTs	BRAMs[1]	Flip-Flops
Recycle buffer	1.1K(0.3%)	8(1.2%)	1K(0.1%)
Decompressor	56K(14.2%)	50(7.0%)	37K(4.7%)
CAPI2 interface	82K(20.8%)	238(33.0%)	79K(10.0%)
Total	138K(35.0%)	288(40.0%)	116K(14.7%)

[1] One 18kb BRAM is counted as a half of one 36kb BRAM.



- 250MHz
- 14% LUTs
- Up to 7.2GB/s
- 10x faster than CPU
- 10x more power efficient

J. Fang e.a. ASAP 2019

Quantum & Computer Engineering

TU Delft

Selected References

H. Peter Hofstee:
Power-Constrained Microprocessor Design. *ICCD 2002: 14-16*

H. Peter Hofstee:
Power Efficient Processor Architecture and The Cell Processor. *HPCA 2005: 258-262*

James A. Kahle, Michael N. Day, H. Peter Hofstee, Charles R. Johns, Theodore R. Maeurer, David J. Shippy:
Introduction to the Cell multiprocessor. IBM J. Res. Dev. 49(4-5): 589-604 (2005)

Michael Gschwind, H. Peter Hofstee, Brian K. Flachs, Martin Hopkins, Yukio Watanabe, Takeshi Yamazaki:
Synergistic Processing in Cell's Multicore Architecture. IEEE Micro 26(2): 10-24 (2006)

Kanna Shimizu, H. Peter Hofstee, John S. Liberty:
Cell Broadband Engine processor vault security architecture. *IBM J. Res. Dev. 51(5): 521-528 (2007)*

H. Peter Hofstee:
Heterogeneous Multi-core Processors: The Cell Broadband Engine. Multicore Processors and Systems 2009: 271-295

H. Peter Hofstee:
Cell Broadband Engine Processor. Encyclopedia of Parallel Computing 2011: 234-241

(look in DBLP for implementation-oriented papers and more recent work on Big Data Systems & accelerators)

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

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.