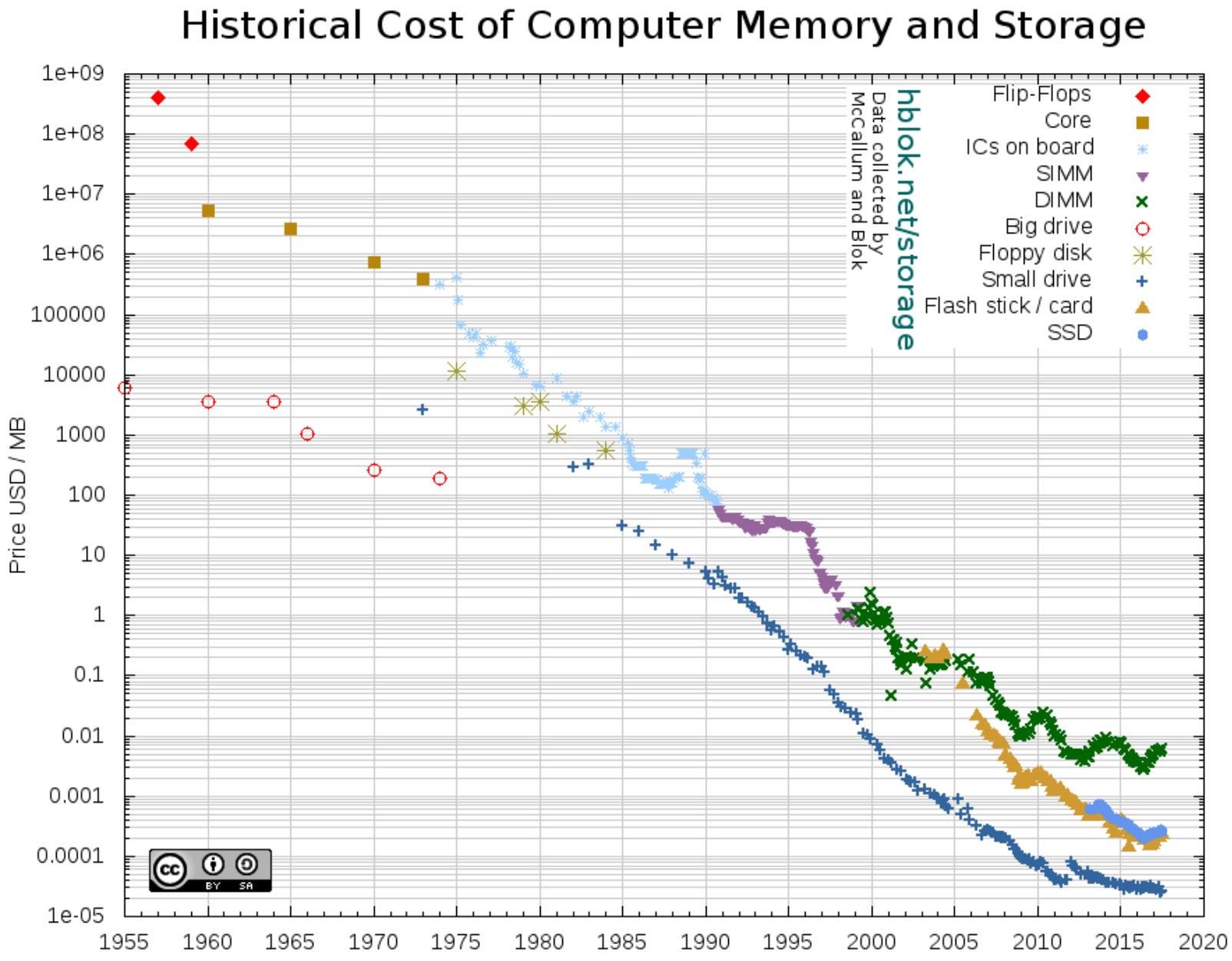
Reimagine Storage Systems for Future Data Centers

Bingsheng He

Dean's Chair Associate Professor

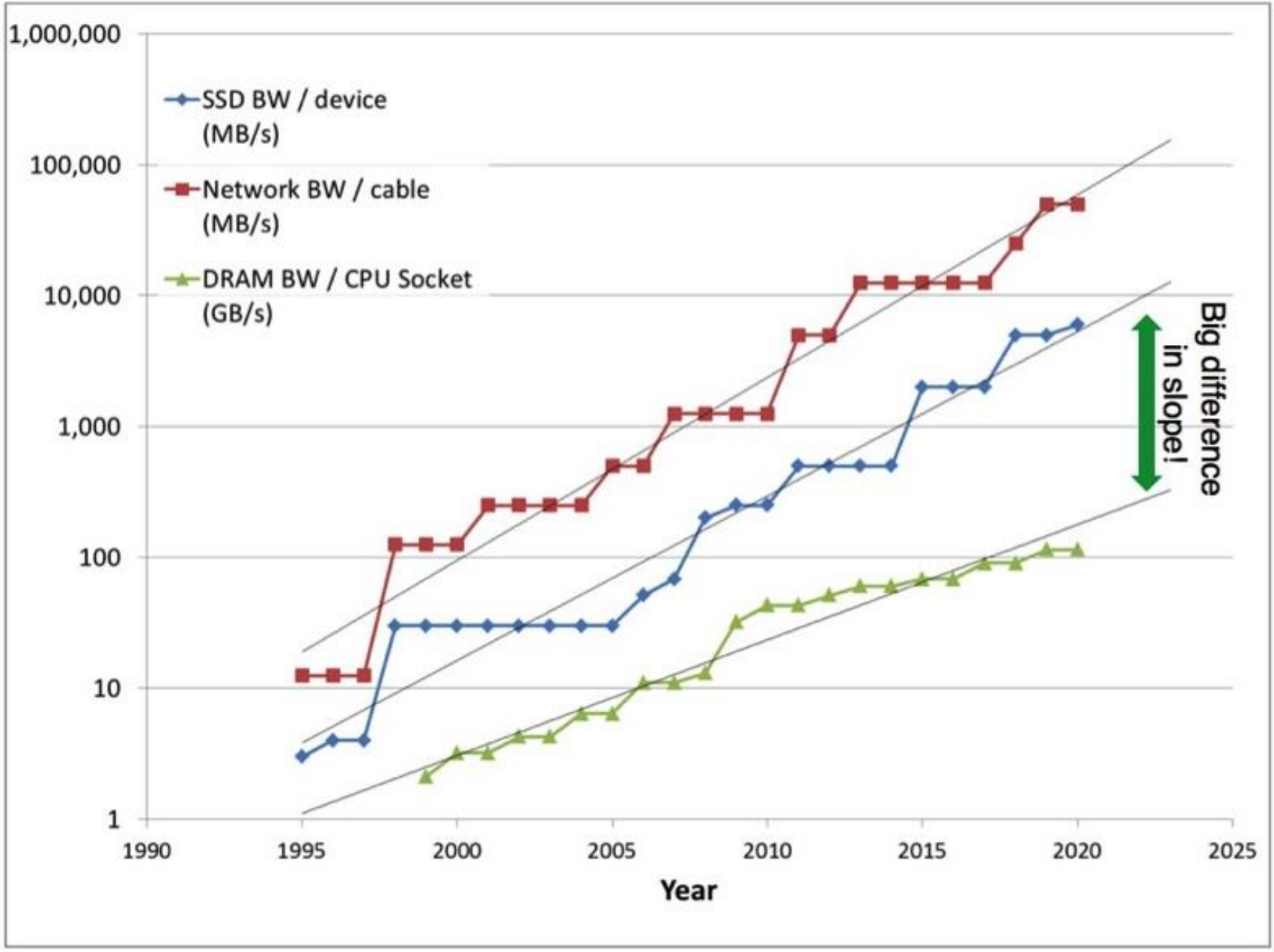
National University of Singapore

Trend of Memory and Storage



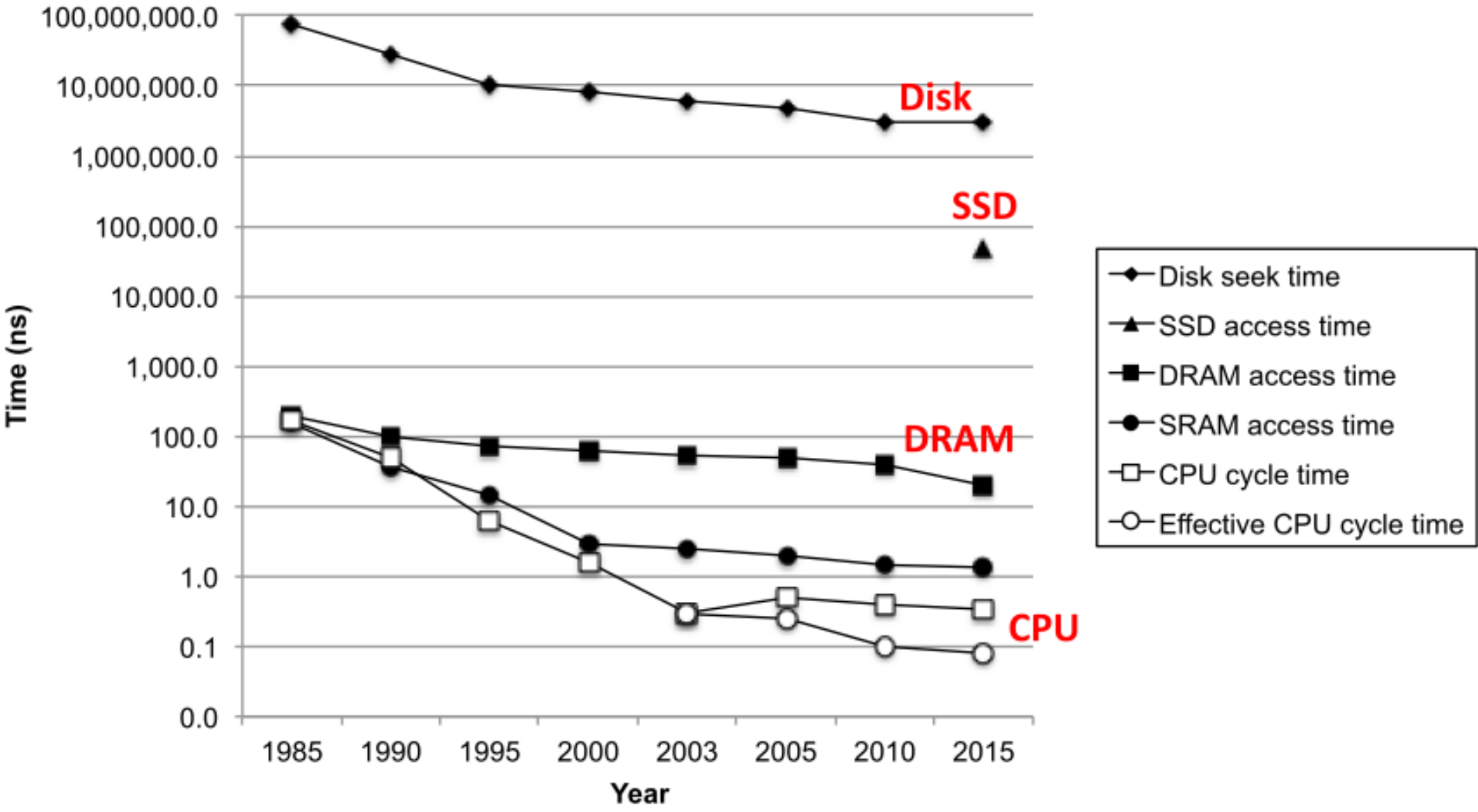
Source: hblok.net/storage

Bandwidth Gap



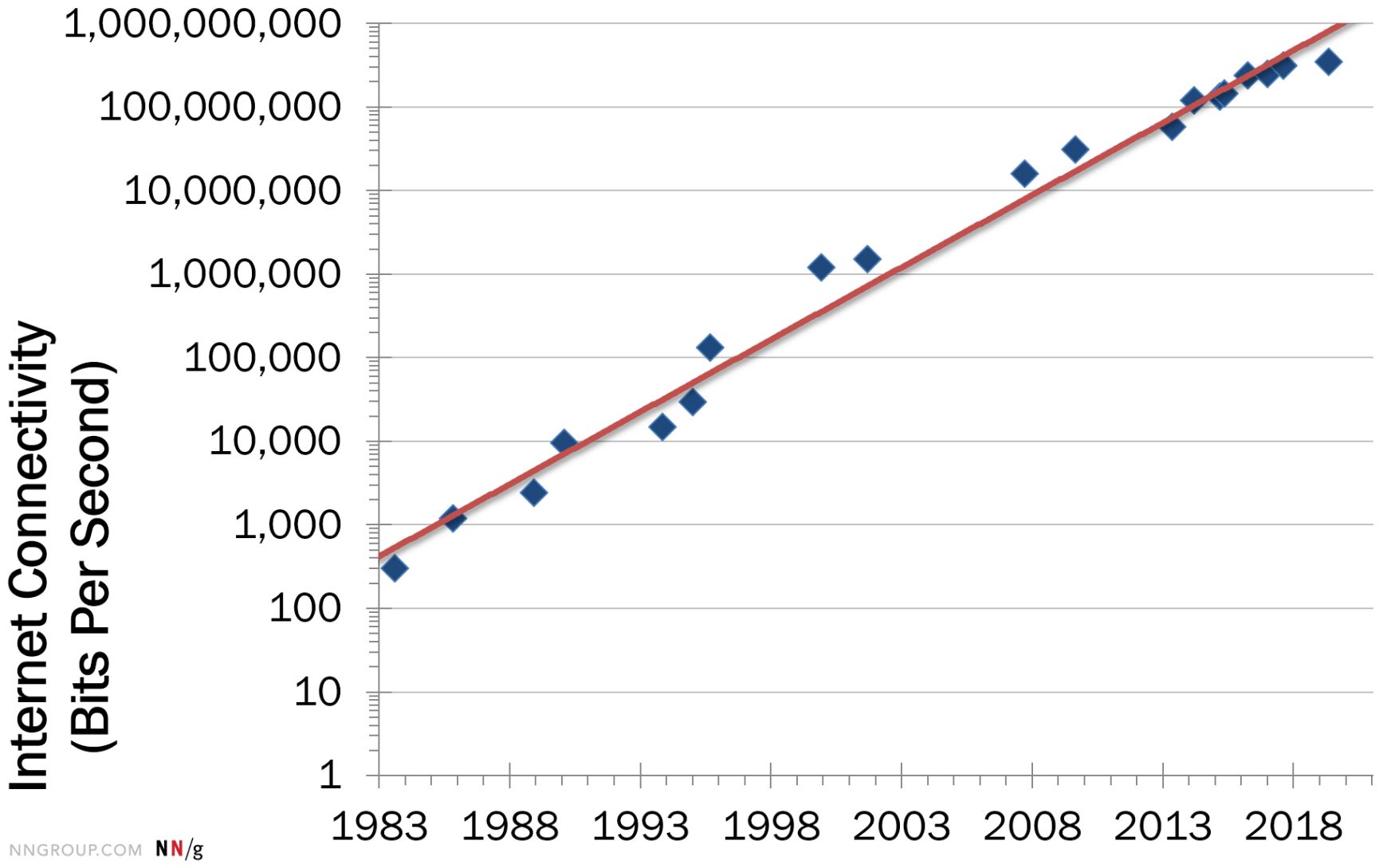
Source: https

Latency Gap



Source: https://www.bowdoin.edu/~sbarker/teaching/courses/systems/18spring/lectures/lec18.pdf/

Trend of network bandwidth



Nielsen's Law of Internet Bandwidth:

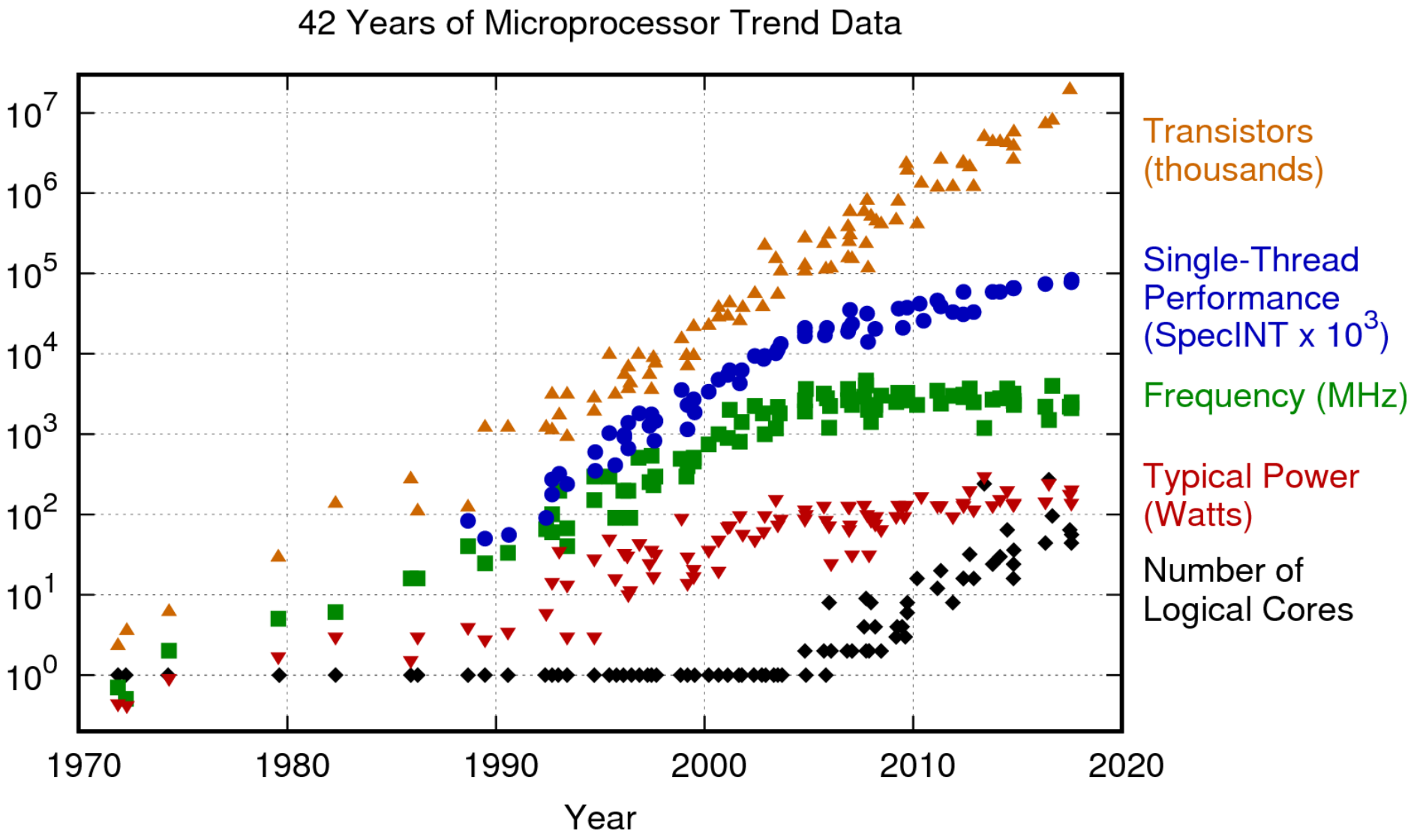
Users' bandwidth grows by 50% per year (10% less than Moore's Law for computer speed).

The new law fits data from 1983 to 2019. https://www.nngroup.com/articles/law-of-bandwidth/

Idea 1: Near-Data Processing Architectures

- **Data movement is costly.**
- **Move computation to the data, rather than moving data to computation.**
- **This is a fundamental shift from the current hardware architecture.**
- **We need new HW-SW codesign.**
  - DPU
  - Smart NIC

End of Moore's Law?

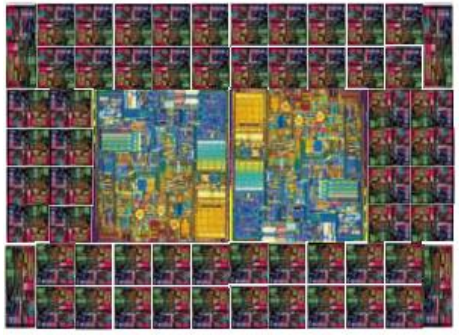


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

https://www.blopig.com/blog/2019/01/making-the-most-of-your-cpus-when-using-python/

Emerging HPC Hardware: Parallelism and Heterogeneity

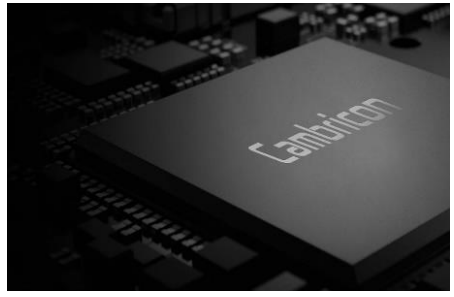
From CPU to accelerators (co-processors)



Many-core CPU

GPU

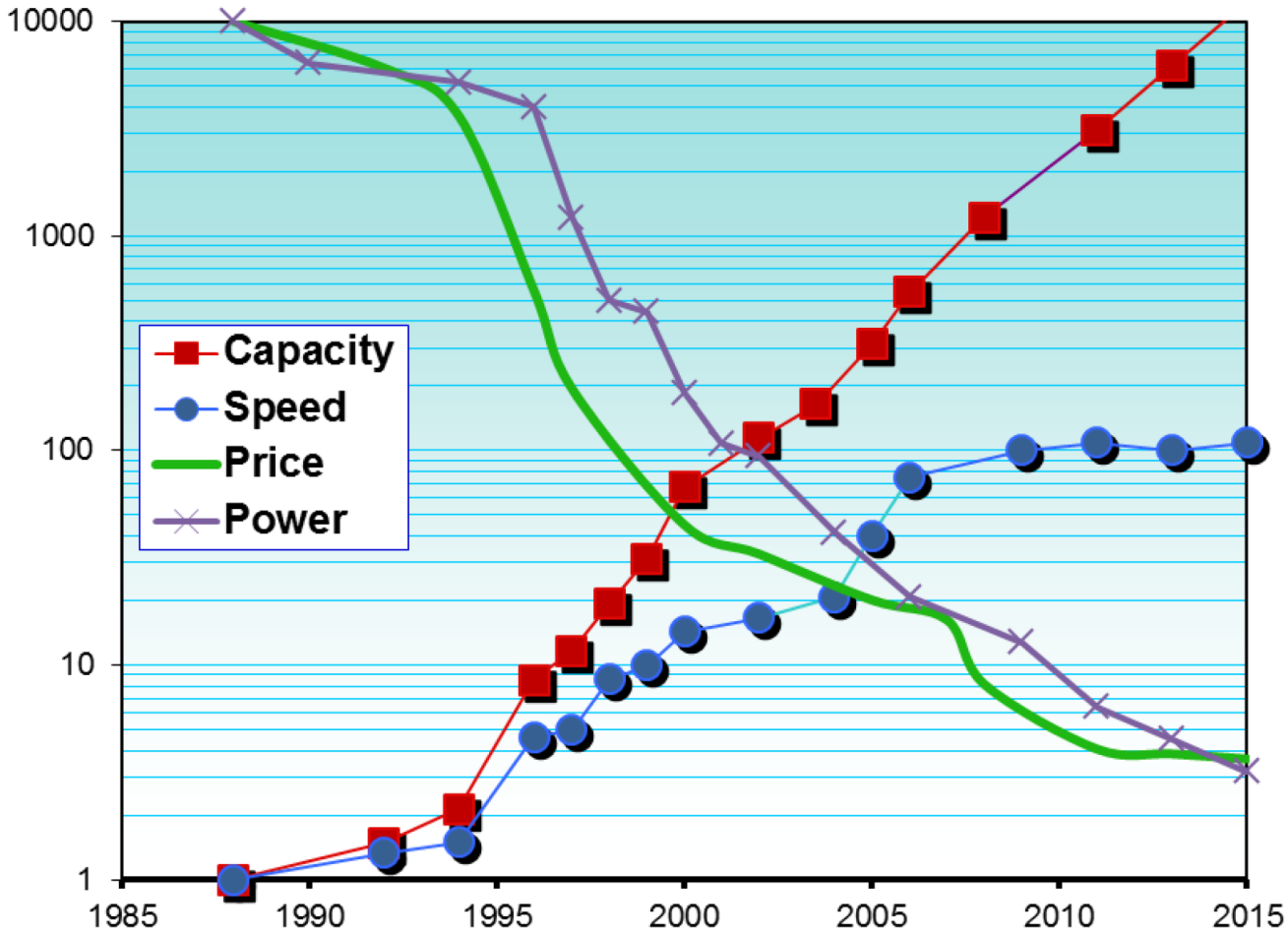
FPGA



ASIC

Figures are adopted from web.

Trend of FPGAs



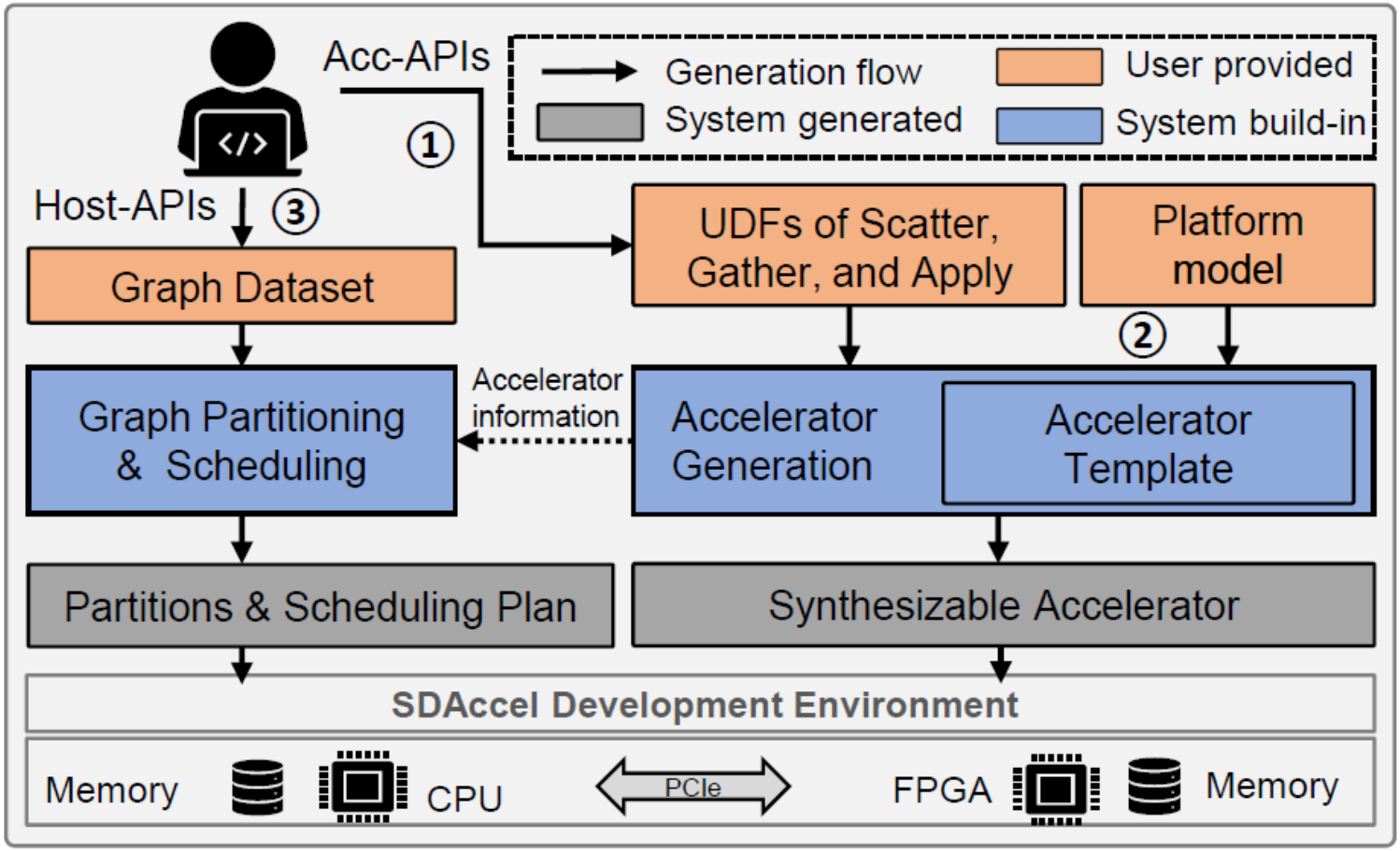
Over the ensuing 30 years:

- FPGA increased in capacity by more than a factor of 10000
- FPGA increased in speed by a factor of 100
- Cost and energy consumption per unit function decreased by more than a factor of 1000

Trimberger, Stephen M. Steve. "Three Ages of FPGAs: A Retrospective on the First Thirty Years of FPGA Technology." *IEEE Solid-State Circuits Magazine, 2018*.

ThunderGP: Fast Graph Processing on HLS-based FPGAs

- **Support in-memory graphs and streaming graphs**
- **Users express their logic with high-level APIs**
- **Fully utilize memory bandwidth**



*More details can be found on https://github.com/Xtra-Computing/ThunderGP/

ThunderGP vs. the State-of-the-art [1]

Algorithm	Dataset	Throughput of work [1]	Our throughput	Speedup
SpMV	WT	1,004	3,138	**3.1x**
SpMV	LJ	1,906	2,860	**1.5x**
PR	R21	3,410	4,759	**1.4x**
PR	LJ	2,110	3,133	**1.5x**
SSSP	WT	2,156	2,954	**1.4x**

Performance (MTEPS) comparison on Xilinx VCU1525 board

*Notes:

1. Since work [1] is not open sourced, performance is collected from their paper.
2. Our performance is measured on-board implementations while their performance is presented with simulation.

[1] Shijie Zhou, Rajgopal Kannan, Viktor K Prasanna, Guna Seetharaman, and Qing Wu. 2019. HitGraph: High-throughput Graph Processing Framework on FPGA. *TPDS* (2019)

ThunderGP is Publicly Available

- **GitHub:** https://github.com/Xtra-Computing/ThunderGP
- **Also featured at** Xilinx Apps and Libraries
- **Invited tutorial at HPCA 2021 with Xilinx**

ThunderGP

ThunderGP: HLS-based Graph Processing Framework on FPGAs

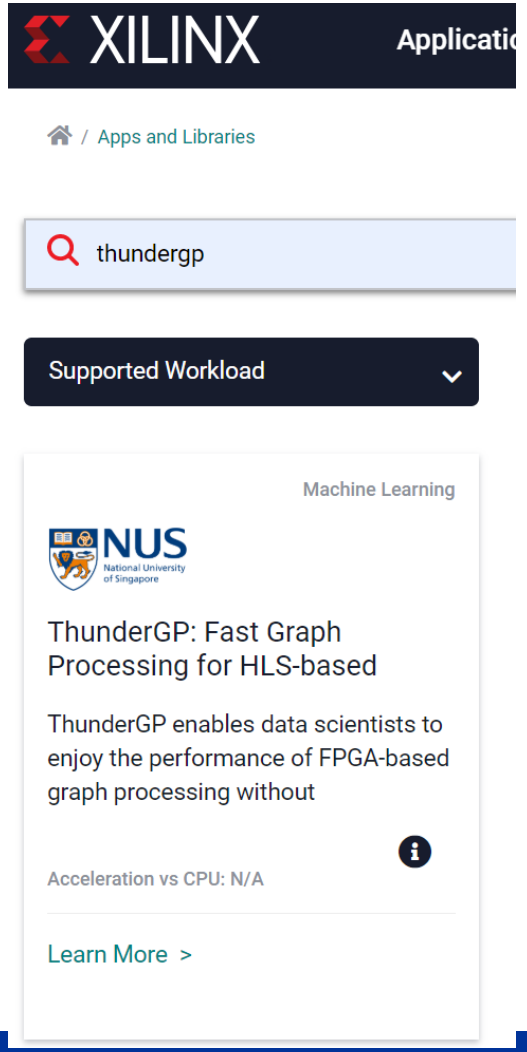
What's new?

ThunderGP won the third place in 2020 Xilinx Adaptive Computing Developer Contest, top 9 out of 79 teams.

ThunderGP is accepted to be FPGA 2021. Read the paper.

ThunderGP is featured at Xilinx Apps and Libraries.

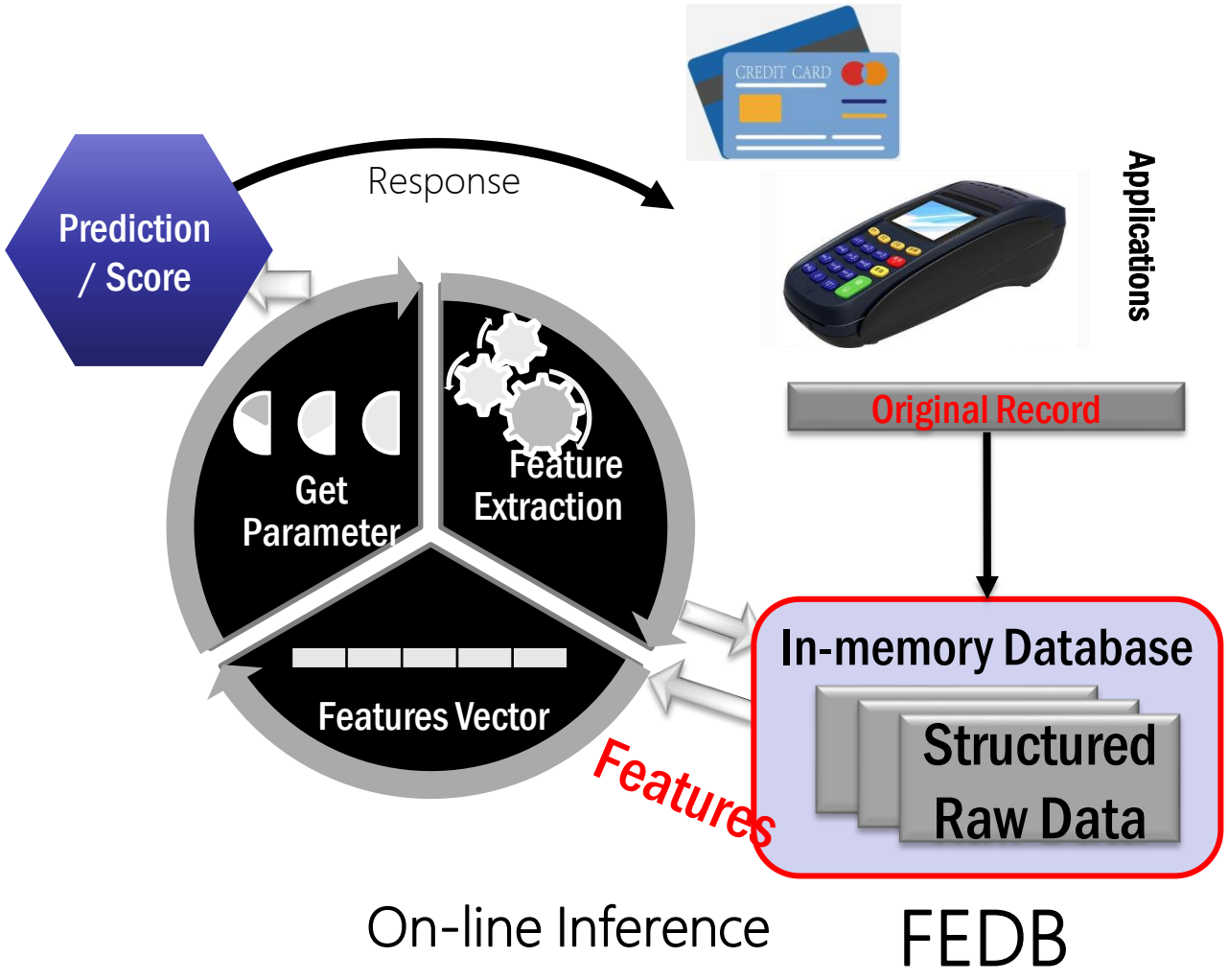
ThunderGP was presented at XACC@NUS Workshop Series 2020: Reconfigurable Computing Systems. see Slides, Video/Youtube, Video/bilibili.



Idea 2: In-Memory Computing

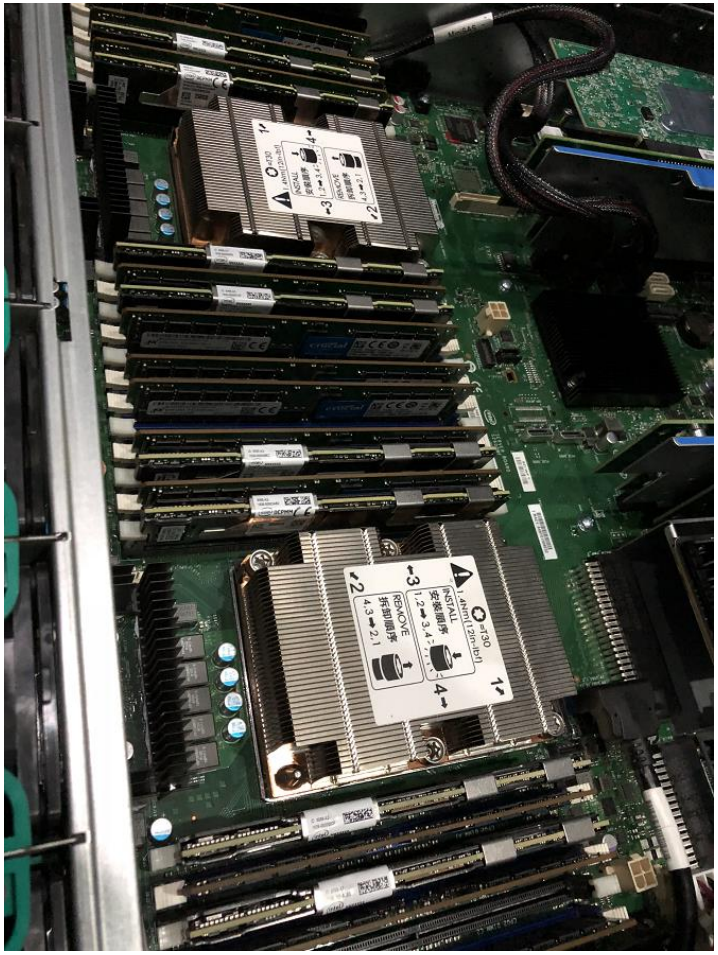
- **Basic idea: fit data into memory for fast accesses.**
- **Pure DRAM solutions only make sense for TB-scale**
  - Cost
  - Energy consumption
- **NVRAM will be the future enable for PB-scale**
- **We need to capture the best of NVRAM (capacity) to achieve close to pure DRAM speed.**

In-Memory Feature Data Management



Joint work with 4paradigm.

Intel® Optane™ DC Persistent Memory (PMEM)

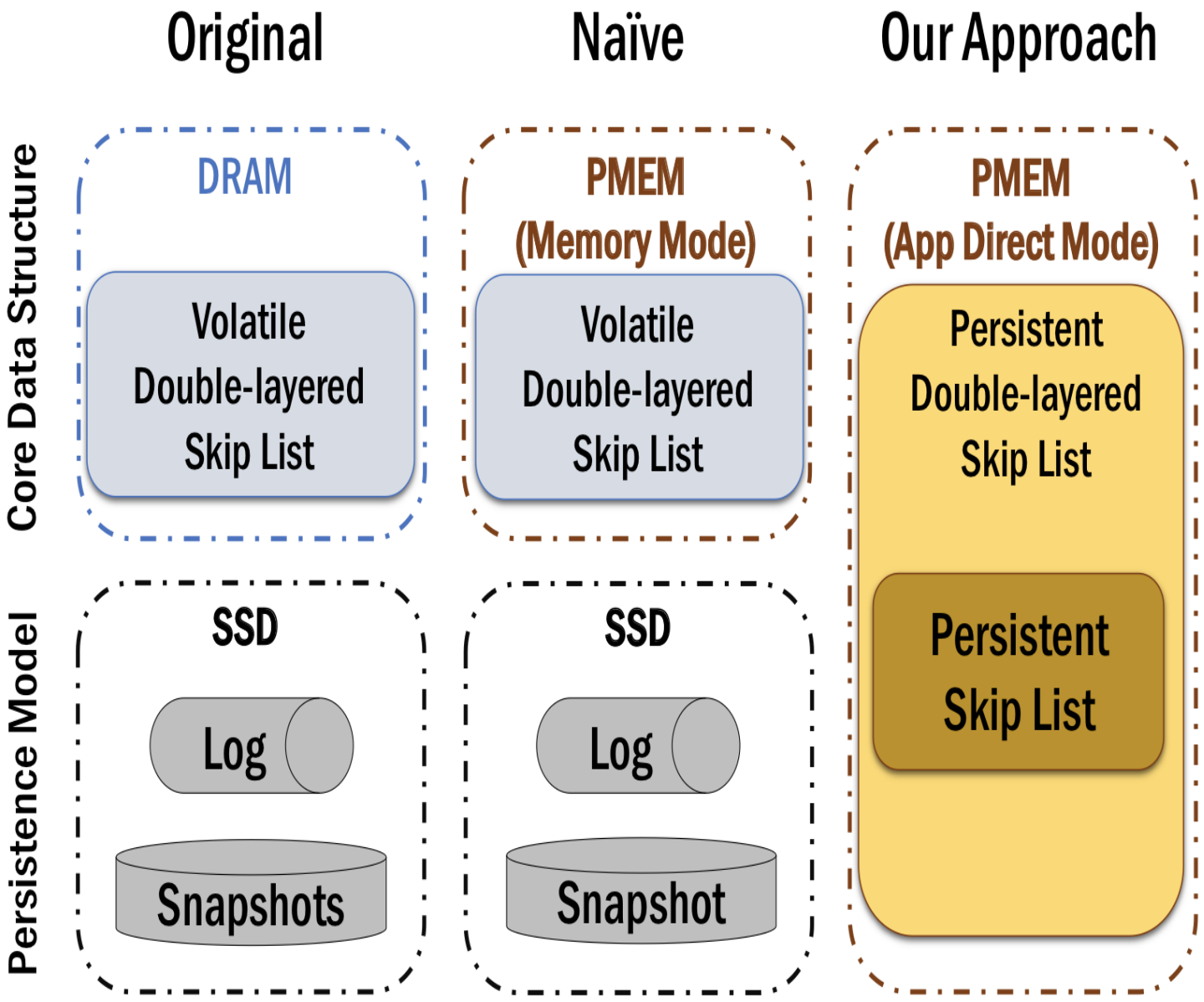


Capacity

DRAM: 4GB ~ 128GB

PMEM: 128GB ~ 512GB

Different Ways of Using PMEM in FEDB



Reduce ~20%
of long tail latency

Save 58.4%
of total cost

Reduce 99.7%
of recovery time

See VLDB 2021: Optimizing In-memory Database Engine for AI-powered On-line Decision Augmentation Using Persistent Memory

Summary

- **Trends of future storage**
- **Idea 1: Near data processing architectures**
- **Idea 2: In-memory computing**

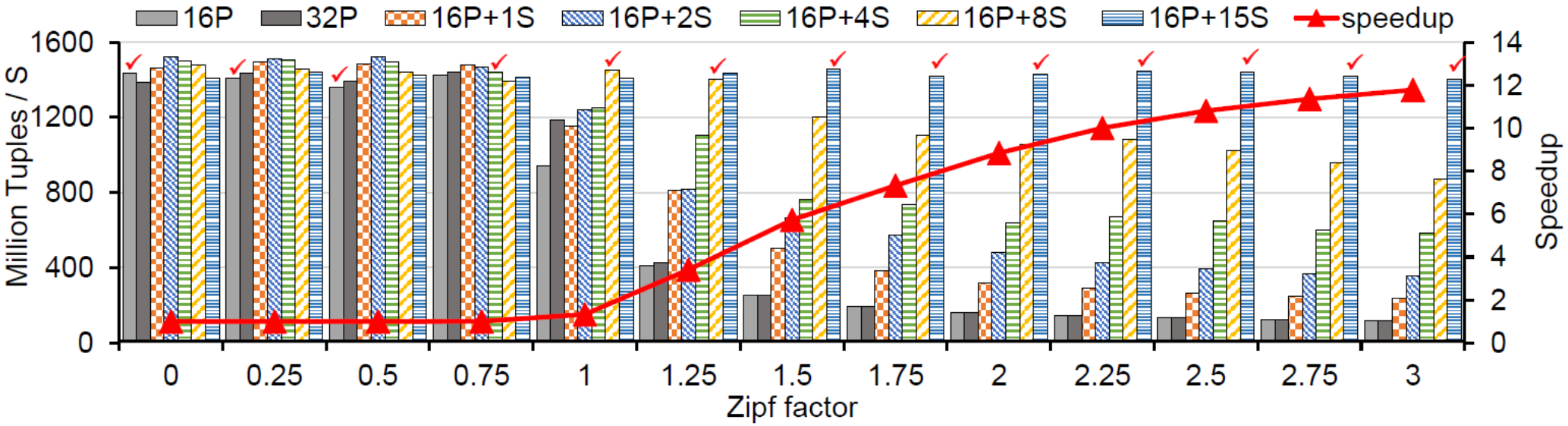
Thank you!

More information:

Dr. Bingsheng He,
hebs@comp.nus.edu.sg
https://www.comp.nus.edu.sg/~hebs/

Skew-Oblivious HLS-based Data Shuffling

- The throughput can be “skew-oblivious”.
- Significant speedup over the baseline that does not have skew handling mechanism.



The throughput of HyperLogLog implementations with different number of SecPEs over Zipf distributions.

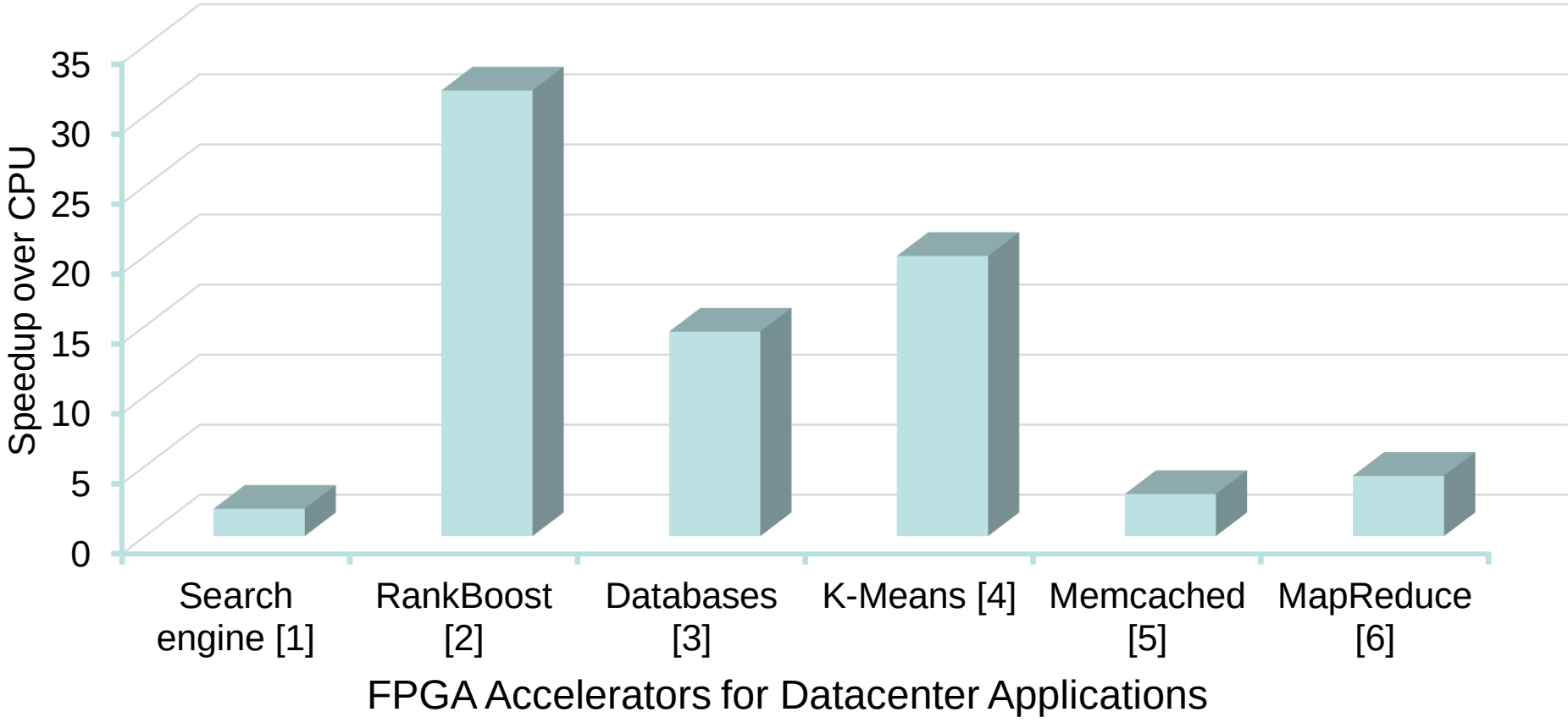
Xinyu Chen, Hongshi Tan, Yao Chen, Bingsheng He, Weng-Fai Wong, Deming Chen:
Skew-Oblivious Data Routing for Data-Intensive Applications on FPGAs with HLS. DAC 2021.

FPGAs Moving to Datacenters

- **Microsoft adopts FPGAs to accelerate Bing search engine [1] and starts BrainWave project [2] with FPGAs.**
- **Baidu takes FPGA approach to accelerate SQL at scale [3].**
- **Cloud venders such as Amazon, Huawei, Tencent and Alibaba are providing FPGA services.**
- **…**

[1] https://blogs.microsoft.com/ai/build-2018-project-brainwave/
[2] https://www.microsoft.com/en-us/research/project/project-catapult/
[3] https://www.nextplatform.com/2016/08/24/baidu-takes-fpga-approach-accelerating-big-sql/

Promising Performance of FPGA Accelerators



[1] Putnam, Andrew, et al. "A reconfigurable fabric for accelerating large-scale datacenter services." ACM SIGARCH Computer Architecture News.
[2] Shan, Yi, et al. "FPMR: MapReduce framework on FPGA." in FPGA, 2010.
[3] Sukhwani, Bharat, et al. "Database analytics: A reconfigurable-computing approach." Micro, 2013.
[4] Choi, Yuk-Ming, and Hayden Kwok-Hay So. "Map-reduce processing of k-means algorithm with FPGA-accelerated computer cluster." in ASAP, 2014.
[5] Blott, Michaela, et al. "Scaling out to a single-node 80Gbps memcached server with 40terabytes of memory." HotStorage, 2015.
[6] Kachris, Christoforos, et al. "An fpga-based integrated mapreduce accelerator platform." Journal of Signal Processing Systems, 2017.