Powering the future of Cloud Service Providers
An inside view from Systems Research

Dr. Robert Haas
Department Head,
Cloud & Computing Infrastructure
IBM Research – Zurich

rha@zurich.ibm.com
https://www.linkedin.com/in/rohaas/
https://twitter.com/robert_r_haas

Please Note

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM’s sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including 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 results similar to those stated here.

Outline

DATA

- Compute
- Move
- Store

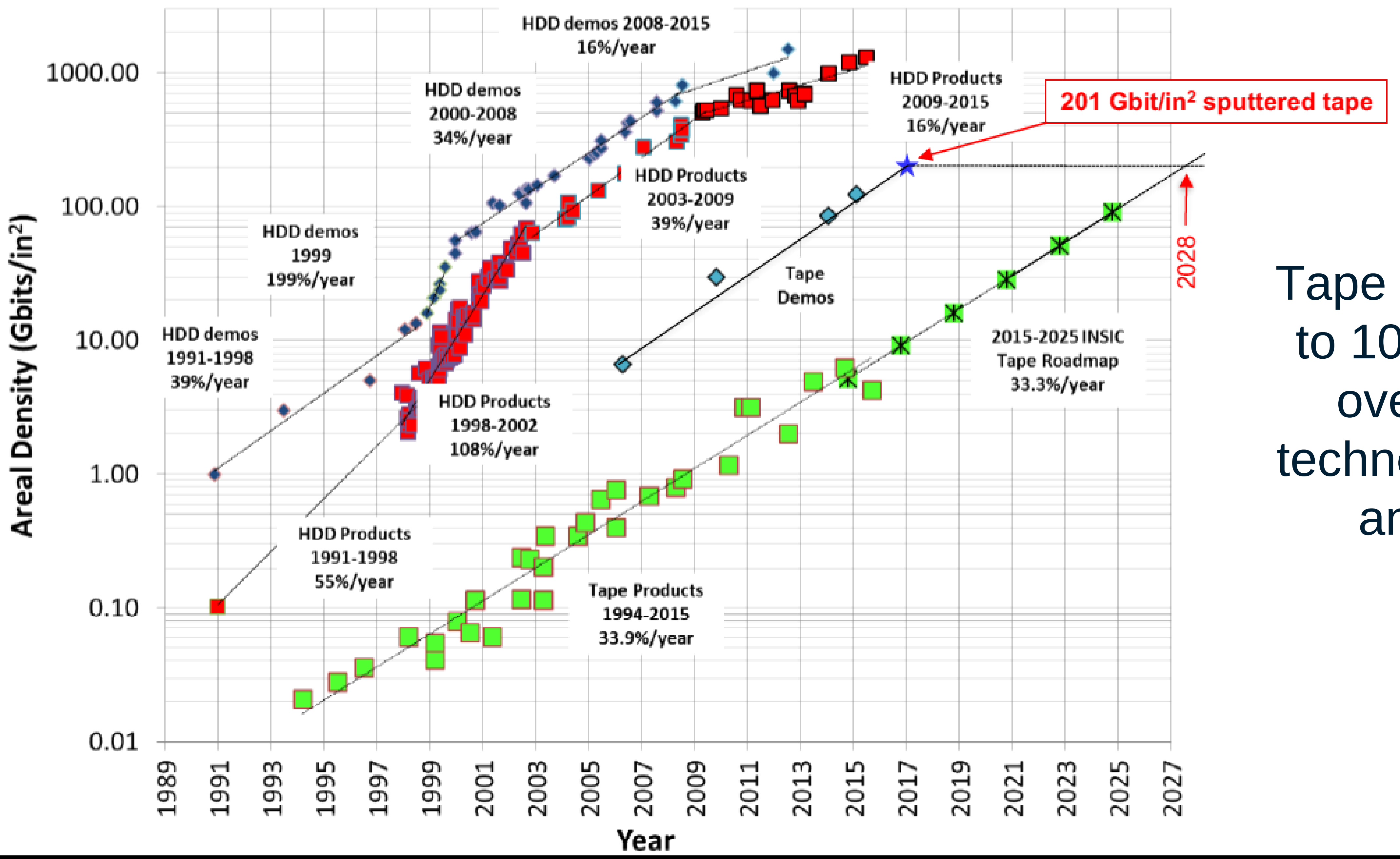
1) Storing Data

Pushing the limits of density in tape and Flash to lower costs

Magnetic Recording Areal Density Trends

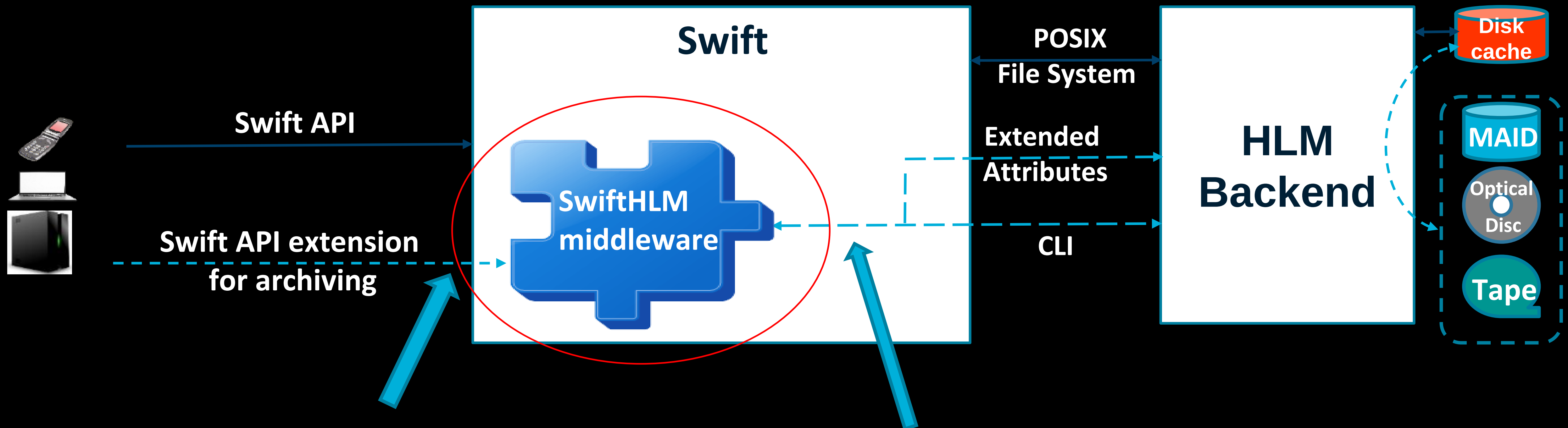
2015: IBM-FujiFilm demonstration of 123 Gb/in^2 on BaFe tape

2017: IBM-Sony demonstration of 201 Gb/in^2 on Sputtered Tape



Tape will maintain its 5x to 10x cost advantage over other storage technologies for at least another decade.

Introducing the Object API for High Latency Media (HLM)



Object API extension for HLM archiving:

- Migrate (Disk -> High-Latency Media, async)
- Recall (High-Latency Media -> Disk, async)
- Query status for Object (sync)
- Query status for Request (sync)
- Object and container level operations

Generic interface for HLM backends, suitable for e.g.:

- IBM Spectrum Archive (LTFS Enterprise Edition)
- IBM Spectrum Protect (TSM/HSM)
- BDT Tape Library Connector (open source)

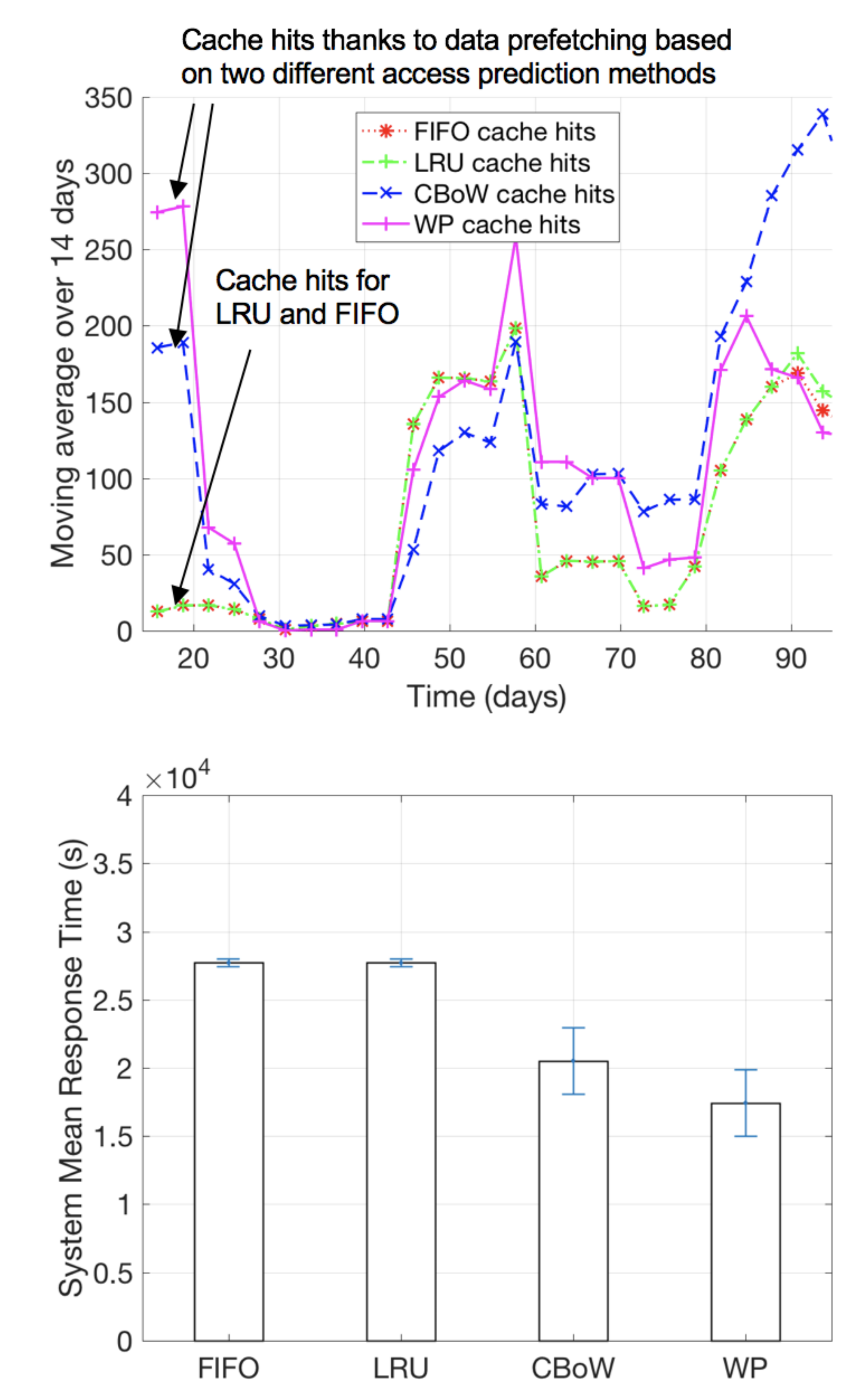
AI for Data Access Prediction

Double cache hit rate and reduce latency by half by predictive pre-fetching data using a disk cache and a tape backend

Evaluated using logs from ASTRON Long Term Archive staging server

Compared to conventional caching algorithms such as FIFO and LRU

- CBoW: conditional bag-of-words model is used when computing meta-data based similarity of files
- WP: word-pair model is used when computing meta-data based similarity of files

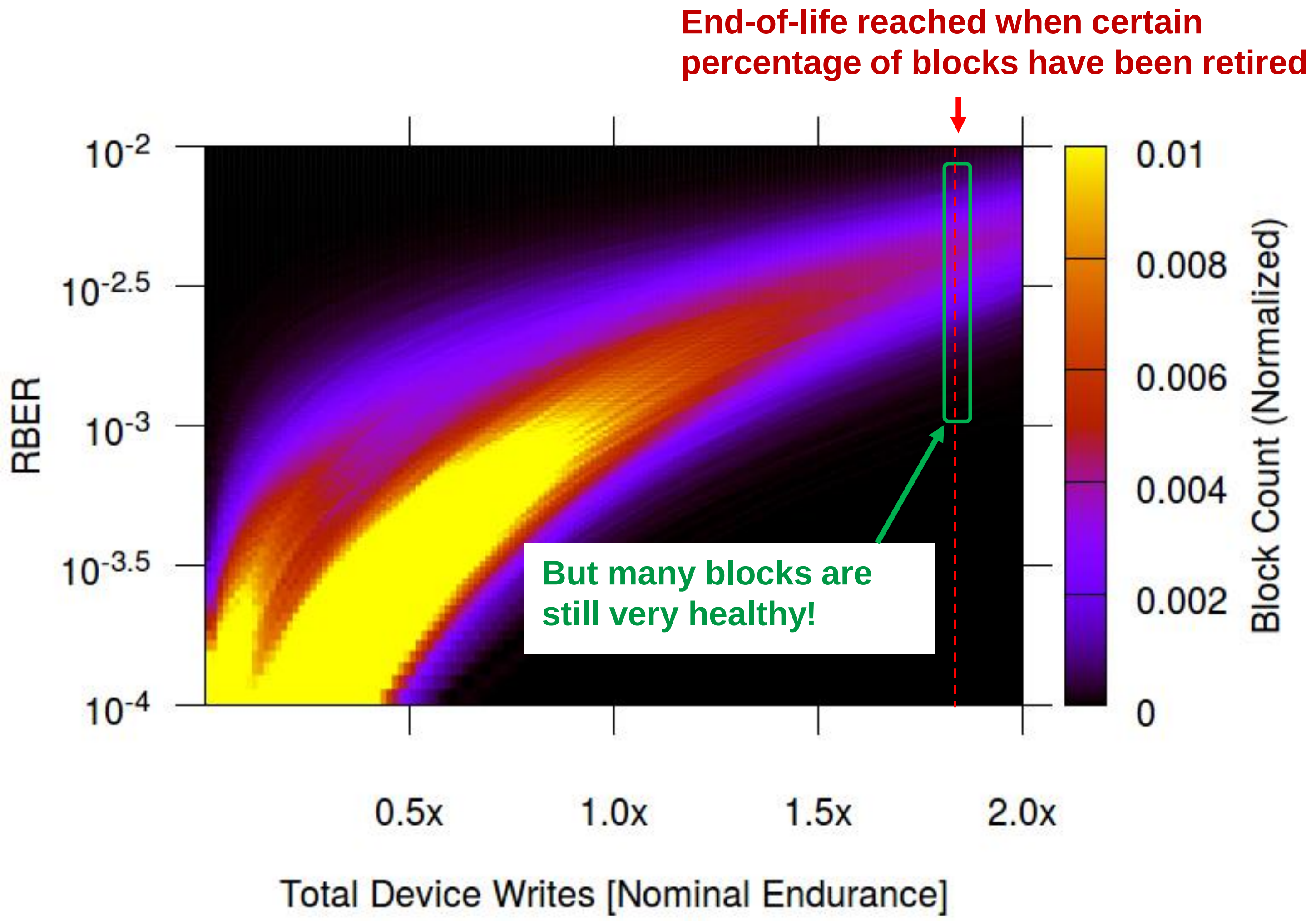


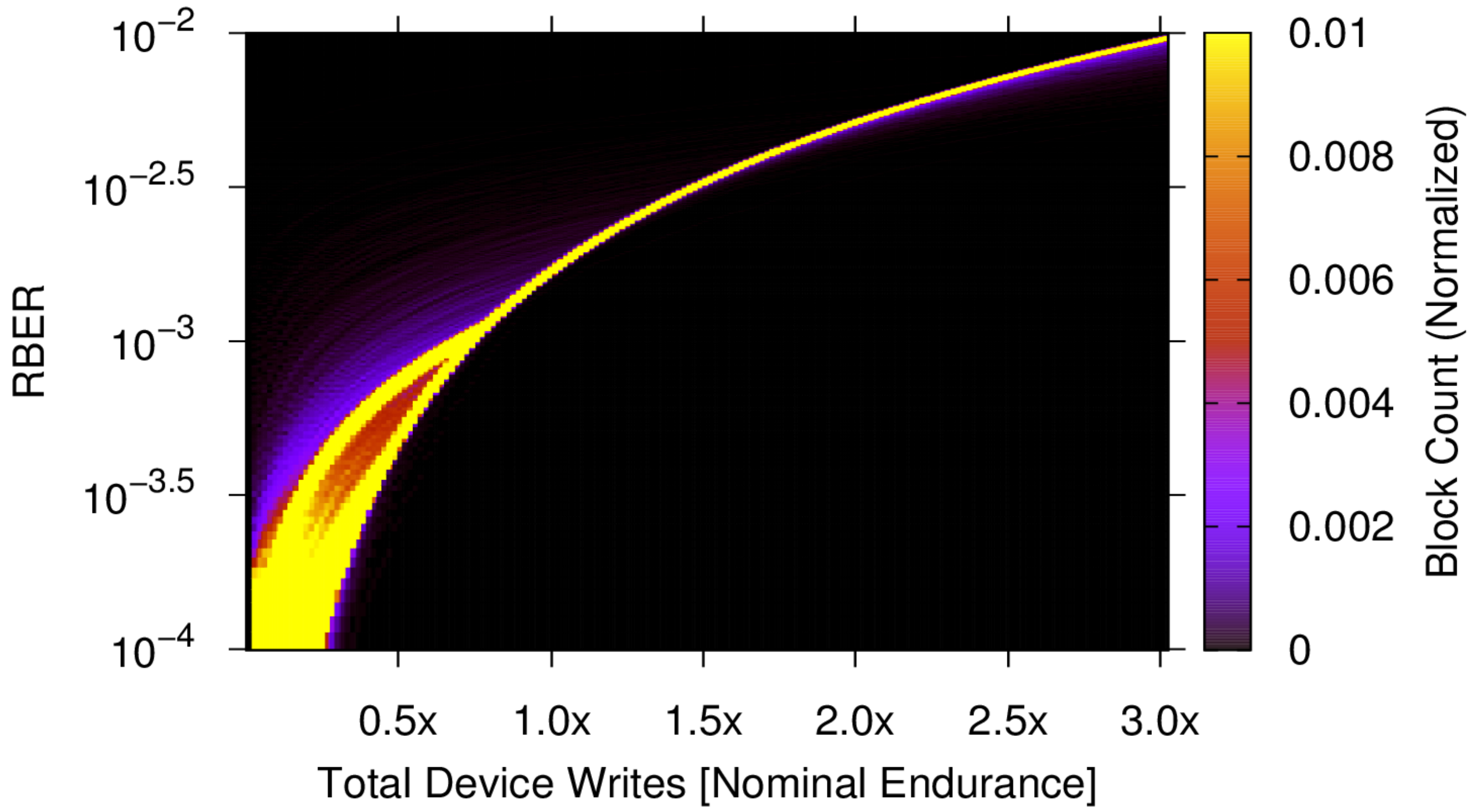
Innovation to Leverage the Densest Flash: 3D-TLC and beyond

Secret Sauce to Squeeze More Write Cycles from Flash

Standard approach
(wear leveling using write-cycle balancing)

Secret Sauce using health binning
and taking advantage of workload skew





2) Moving Data

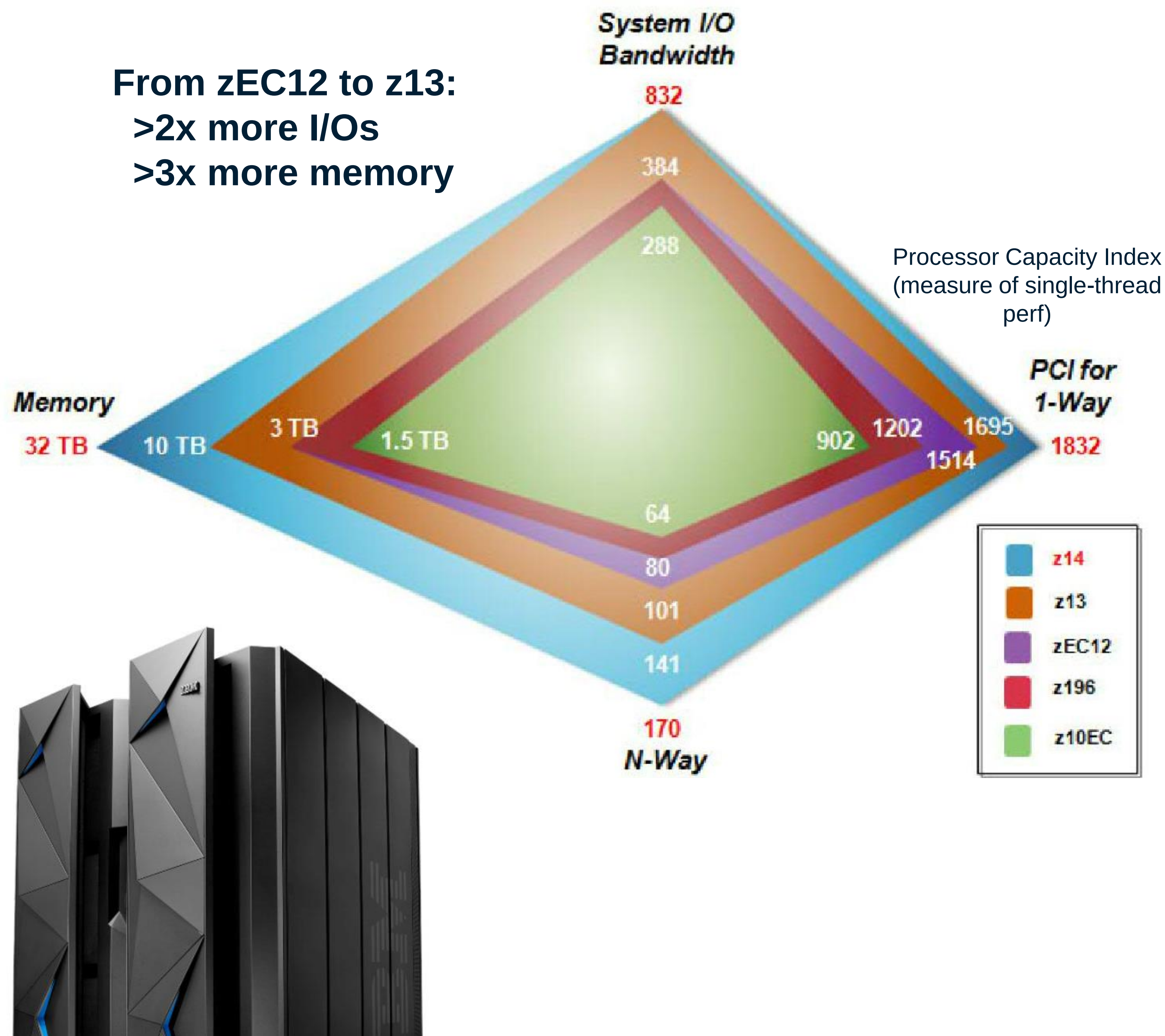
Pushing the limits in terms of speed and lowering power needs,
and letting applications profit from advances in networking and storage performance

Fast and Low-power I/O Links to Feed Data-hungry CPUs and GPUs

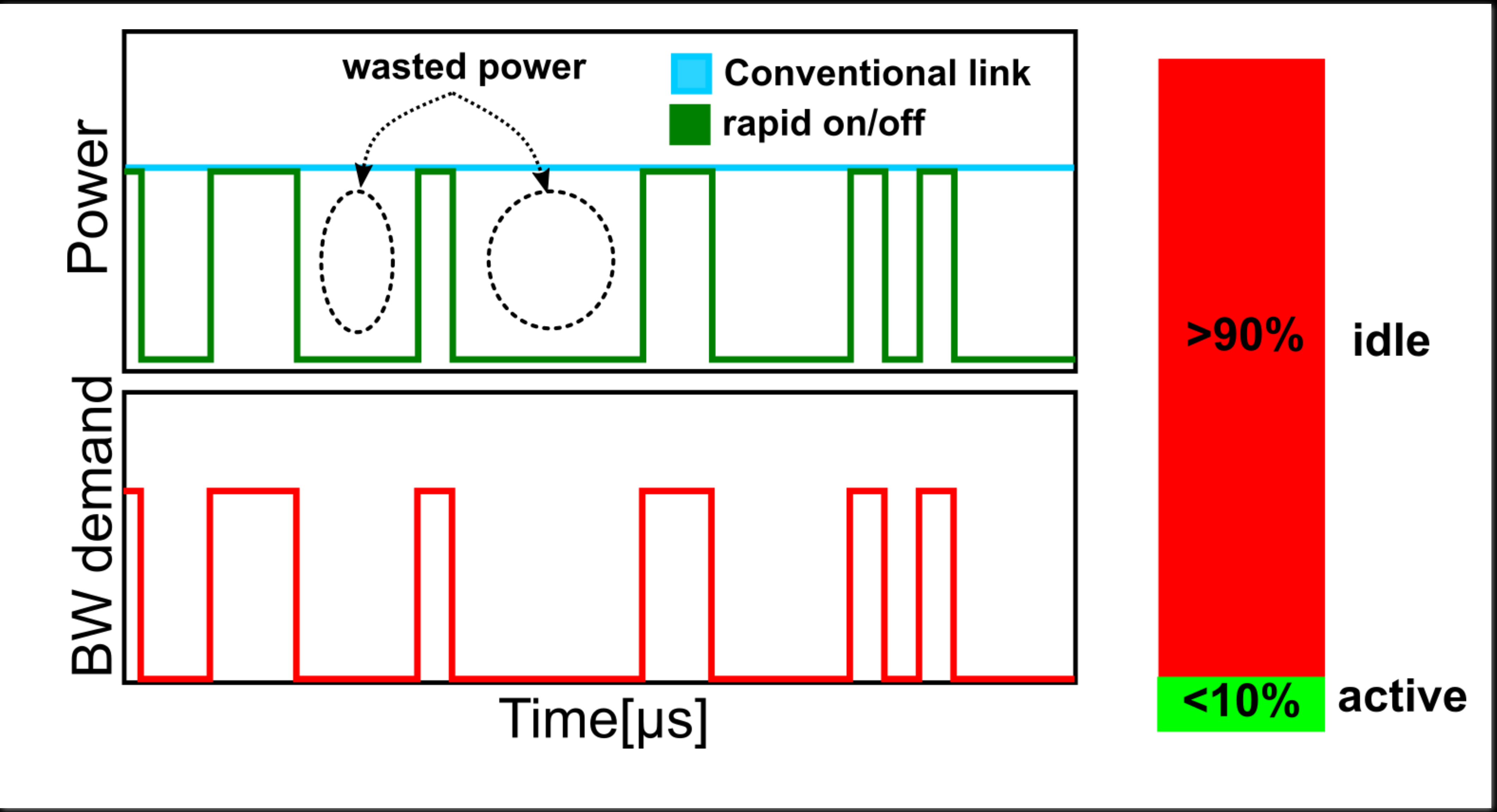
Growing discrepancy between computation and communication
32bit MULT operation = 1pJ
transferring 64bits = 320 pJ

More cores → Need more I/O bandwidth at lower power to fit into thermal envelope

→ I/O links are key enabler for system performance



Wasted Power in Idle Links

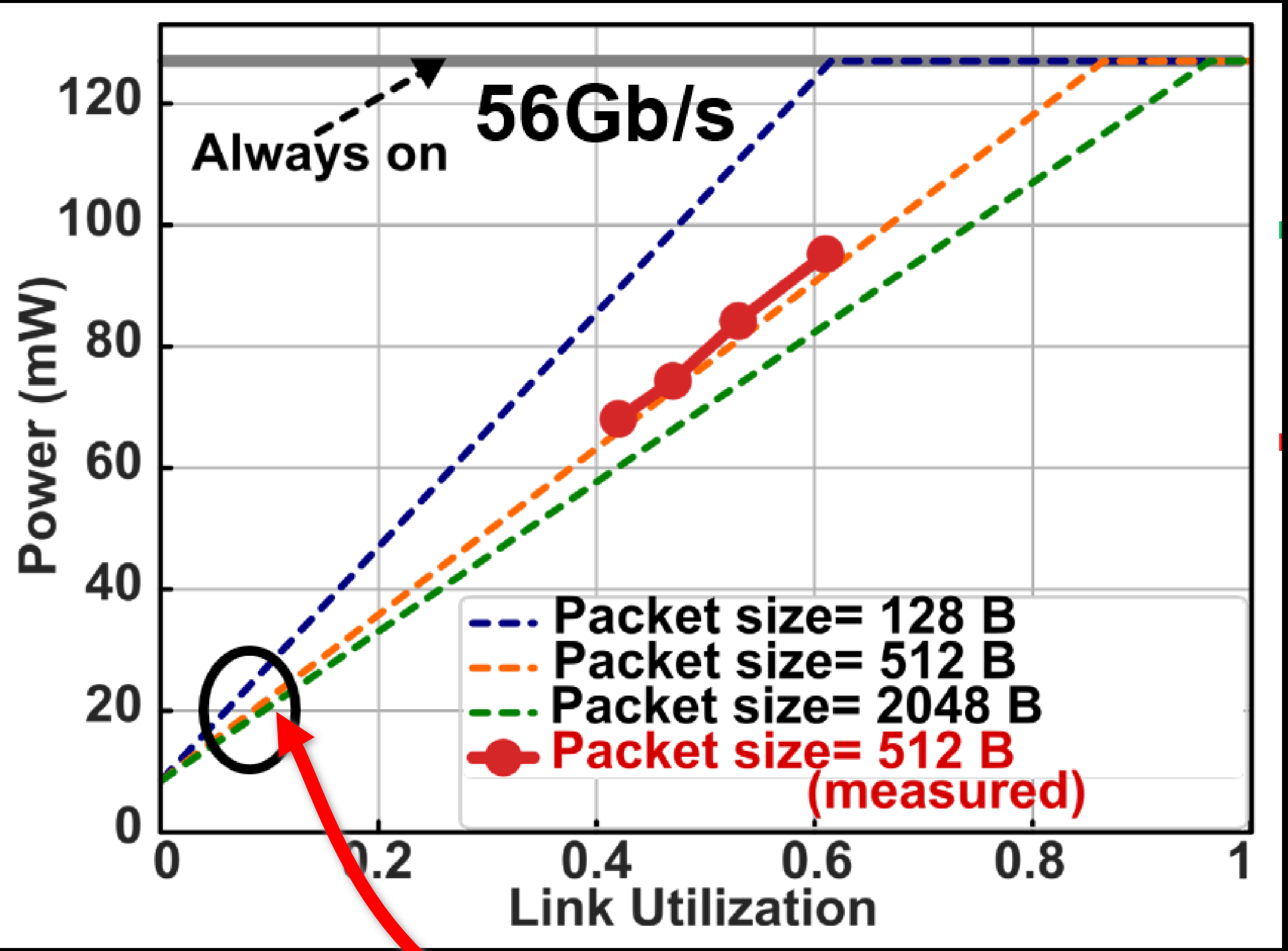


Link utilization in a large date center <10% [1]: idle links burn power.

Rapid on/off: Power scales with link utilization. Enables more bandwidth per chip at fixed Thermal Dissipation Power

[1] A.Roy et al, Sigcomm 2015

Rapid Power-On Links Leads to 85% Less Power Consumption

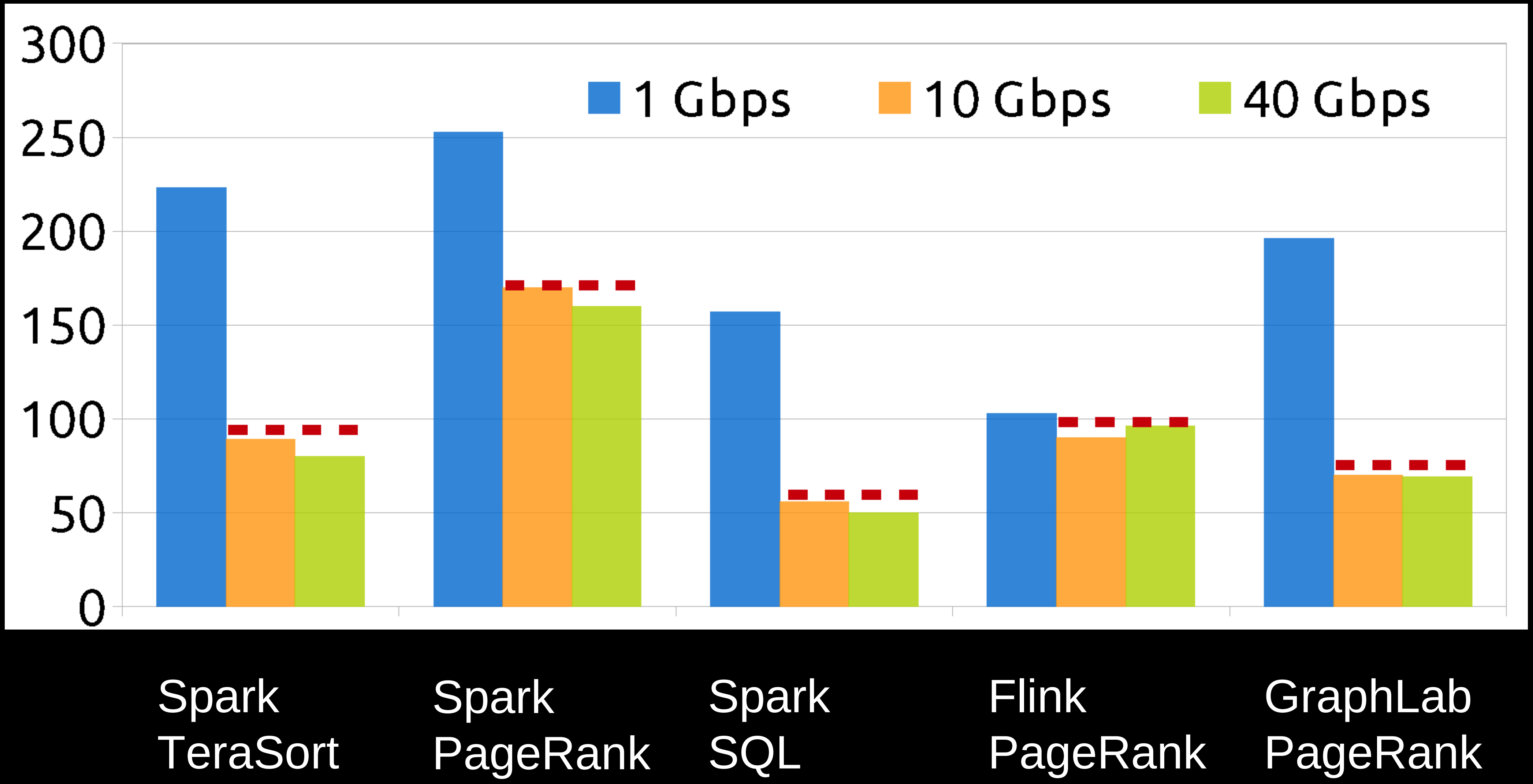


ON power: 128mW (2.1pJ/b) @ 60Gb/s

OFF power: 8mW

At 10% link utilization: 20mW with only 7ns to power on!

Analytics Systems Challenged by High-Performance Networks



Trivedi et al. “On The [Ir]relevance of Network Performance for Data Processing”, Usenix HotCloud 2016

Hardware Performance Does Not Trickle Up the Stack

Applications

“Traditional” Batch Analytics based on MapReduce

- IBM BigInsights
- Hadoop MR

Next-generation Analytics
Interactive, Streaming, Graph, ML

- Spark
- Shark
- GraphX
- MLLib

Hadoop FS (HDFS) API

System Software

Distributed File Storage

Java Runtime

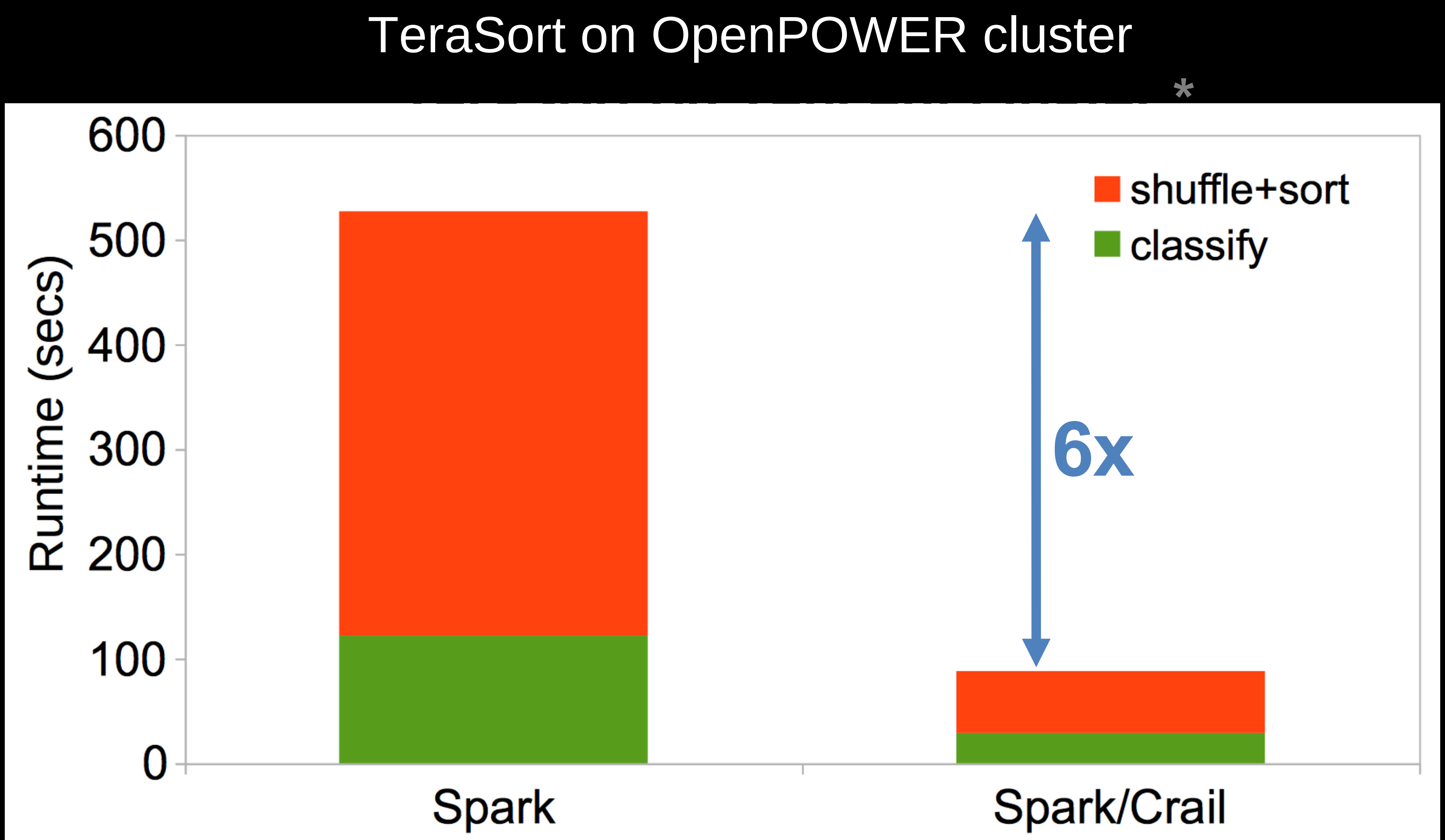
Operating System

Hypervisor

Only 2x better application performance!

HW: 1000x higher throughput, lower latency!

TeraSort on Spark



* 12.8TB sorted on 128 Power8 nodes with 100Gbit Ethernet networking

OpenPOWER cluster
- 128 nodes

Node configuration
- 2 x IBM POWER8 10-core @ 2.9 Ghz
- DRAM: 512GB DDR4
- 4 x 1.2 TB NVMe SSD
- 100GbE Mellanox ConnectX-4 EN (RoCE)

Software
- Ubuntu 16.04 (kernel 4.4.0-31)
- Spark 2.0.0

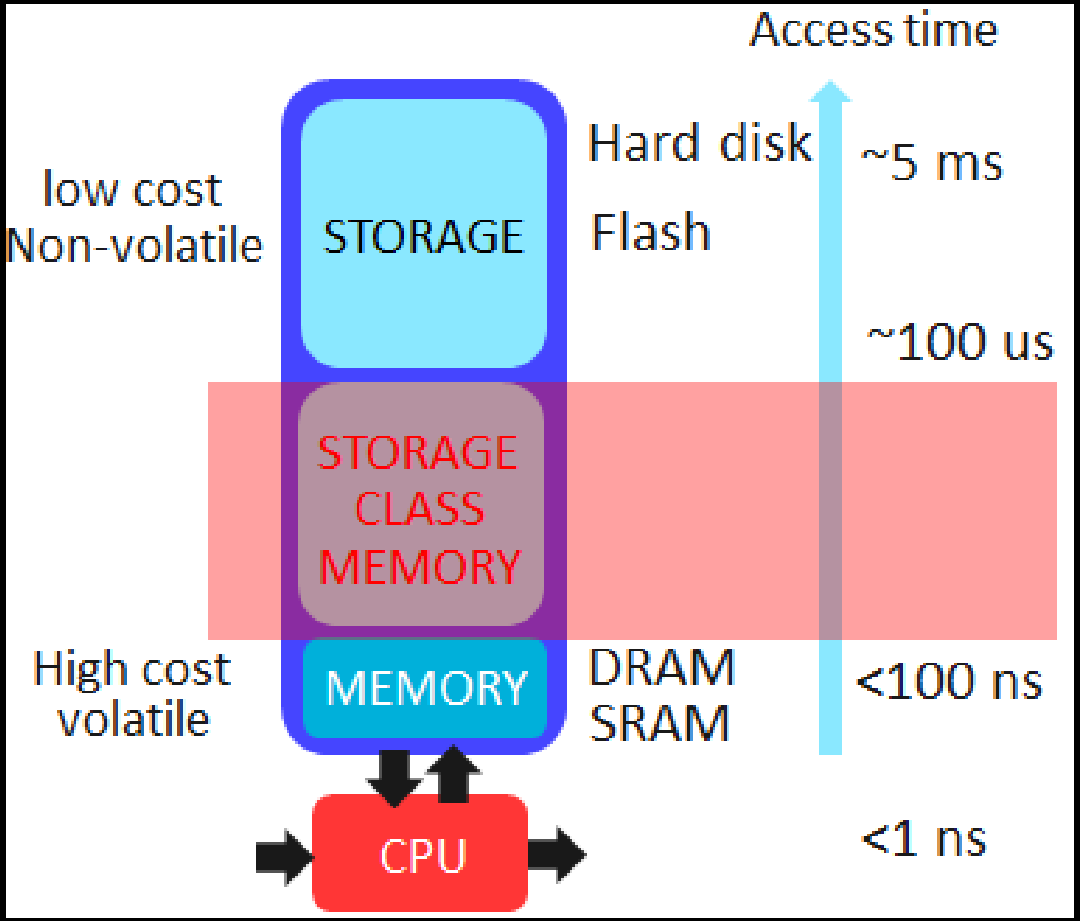
Find out more at http://crail.io/blog

3) Computing on Data

Revolutionize systems with in-place computing

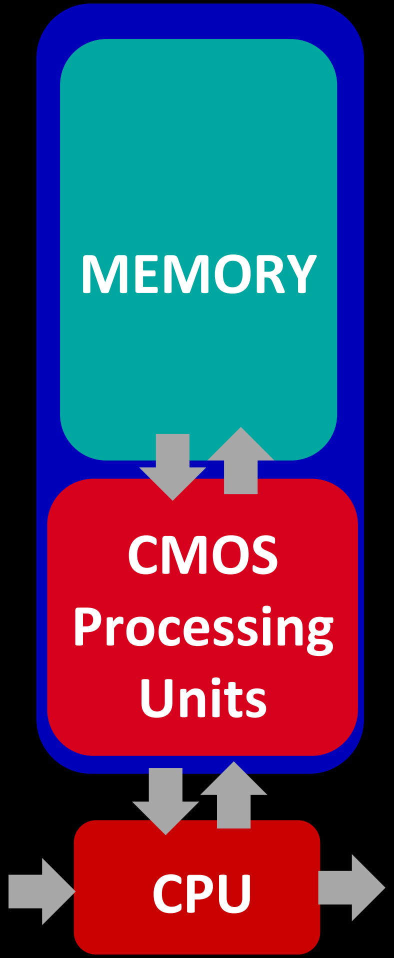
Business As Usual: Continue to Improve von-Neumann Computing

Storage-class memory



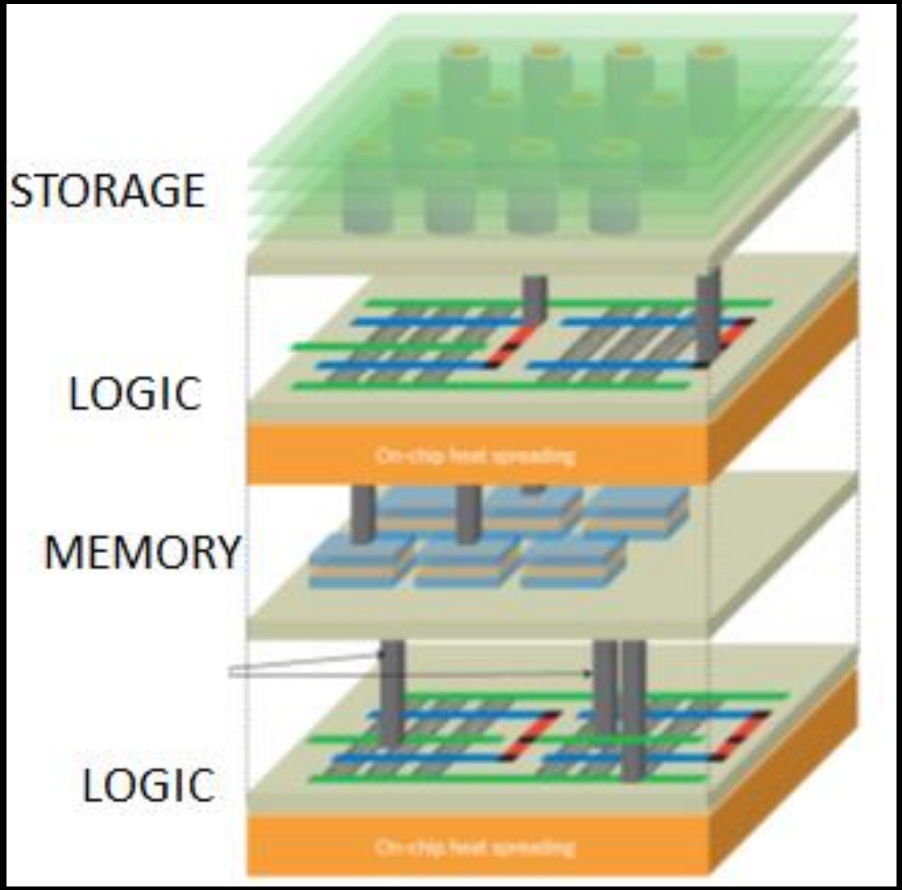
Burr et al., IBM J. Res. Dev., 2008

Near-memory computing



Vermij et al., Proc. ACM CF, 2016

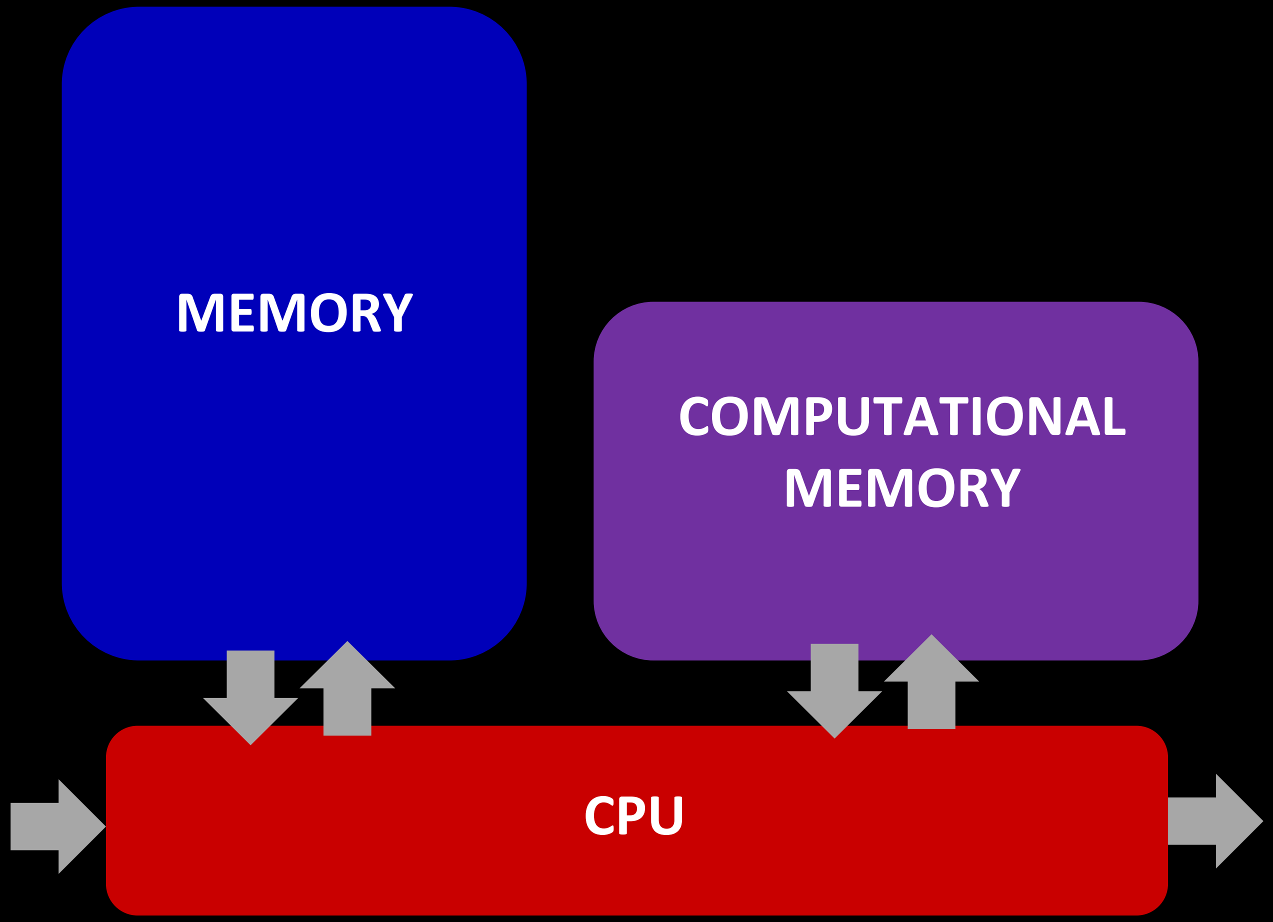
Monolithic 3D integration



Wong, Salahuddin, Nature Nano., 2015

Minimize the time and distance to memory access

Introducing Game-Changing Computational Memory



Perform “certain” computational tasks in place in memory
Not only stores data but performs some calculations on the data

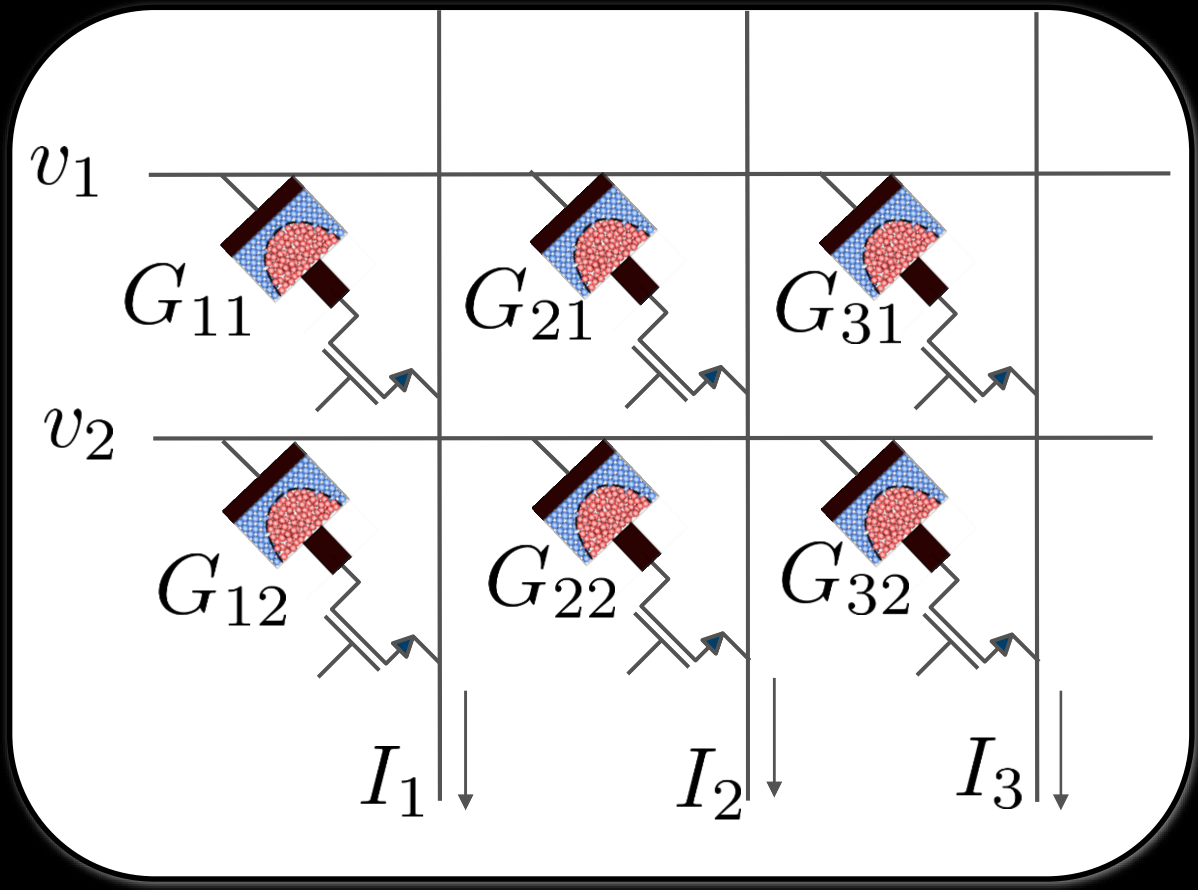
Doing Matrix-vector Multiplication in Computational Memory

$$\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \\ M_{31} & M_{32} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

MAP to conductance values

MAP to read voltage

DECIPHER from the current

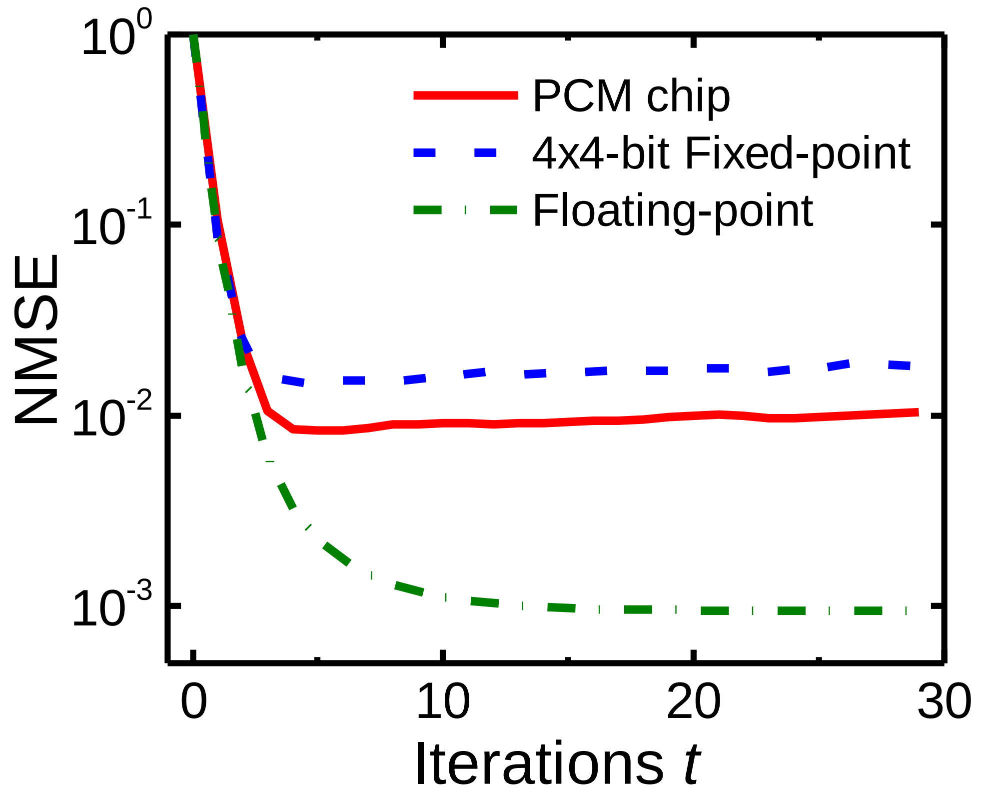


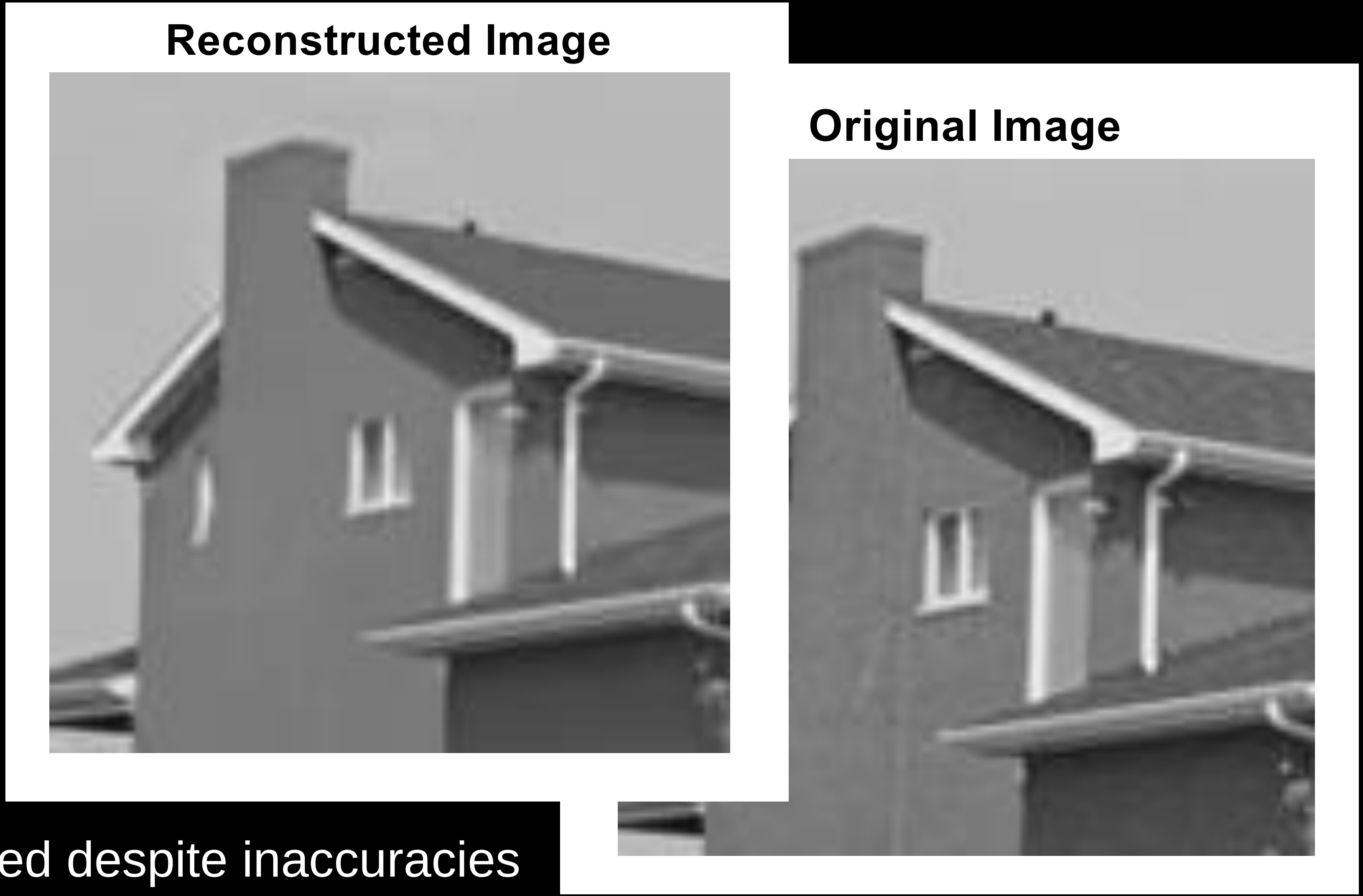
By arranging the PCM devices in a cross-bar configuration one can perform matrix-vector operation with O(1) complexity

Exploits multi-level storage capability and Kirchhoff's circuits laws

Image Reconstruction Experimental Results

128x128 image, 50% sampling rate,
Comp. mem. unit w/ 131,072 PCM devices



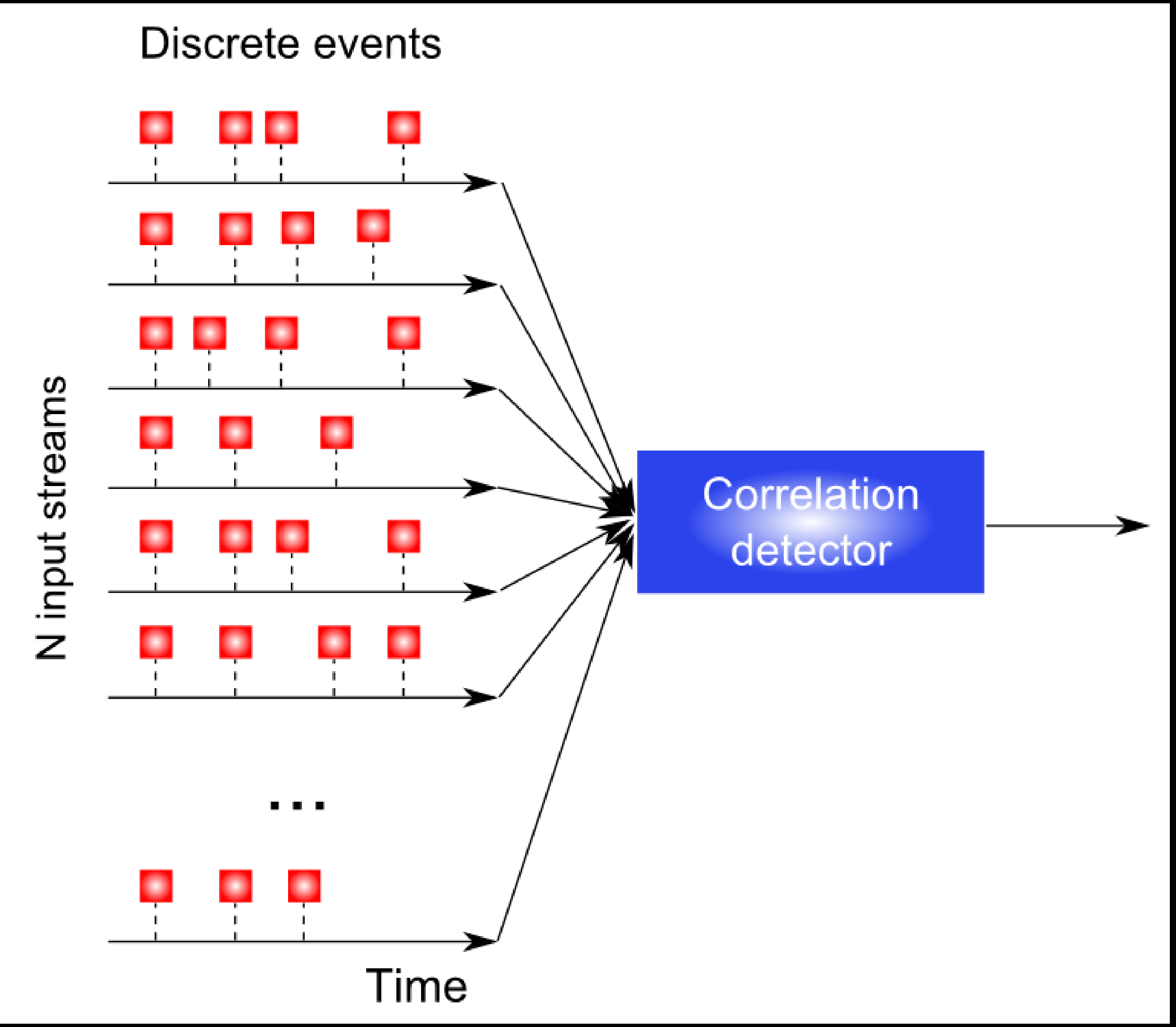


Reasonable reconstruction accuracy achieved despite inaccuracies

Estimated power reduction of 50x compared to using an optimized 4-bit FPGA matrix-vector multiplier that delivers same reconstruction accuracy at same speed

Le Gallo, Sebastian, Cherubini, Giefers, Eleftheriou, Proc. IEDM, 2017

Temporal Correlation Detection



Determine whether some of the input data streams are statistically correlated
Use only unsupervised learning & consume very low power

FINANCE

SCIENCE

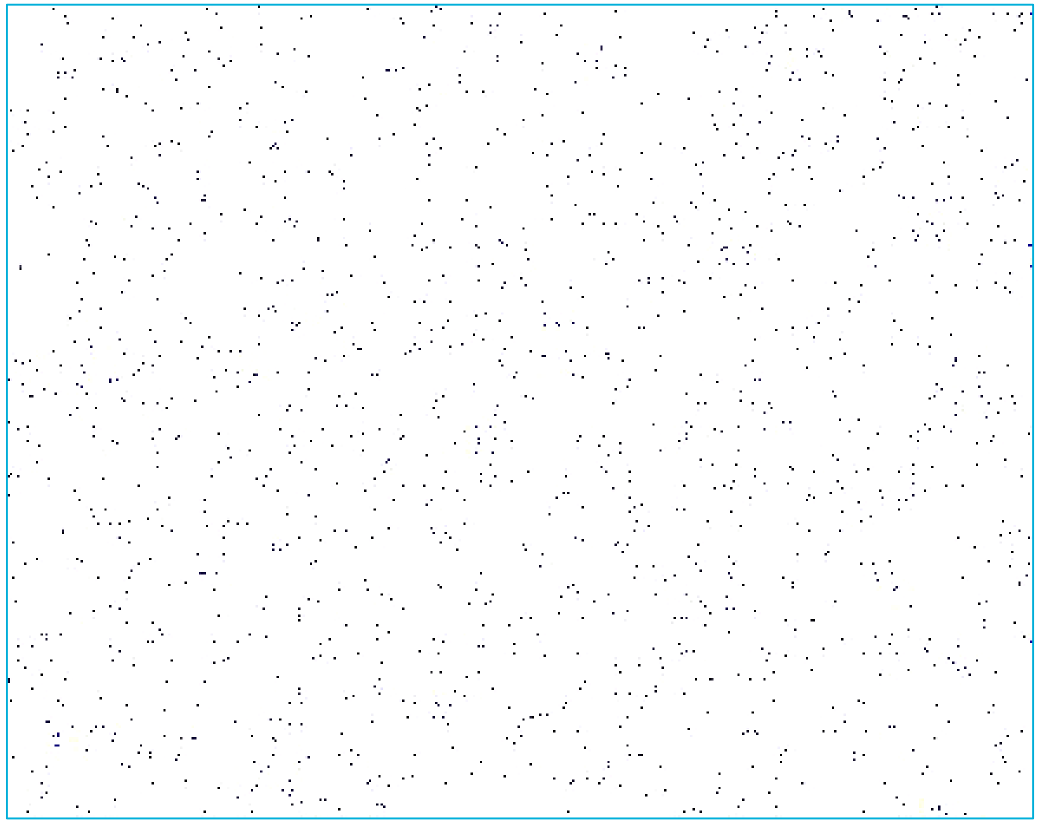
MEDICINE

BIG DATA

...AND MORE

Experimental Results with 1 Million PCM devices

Processes



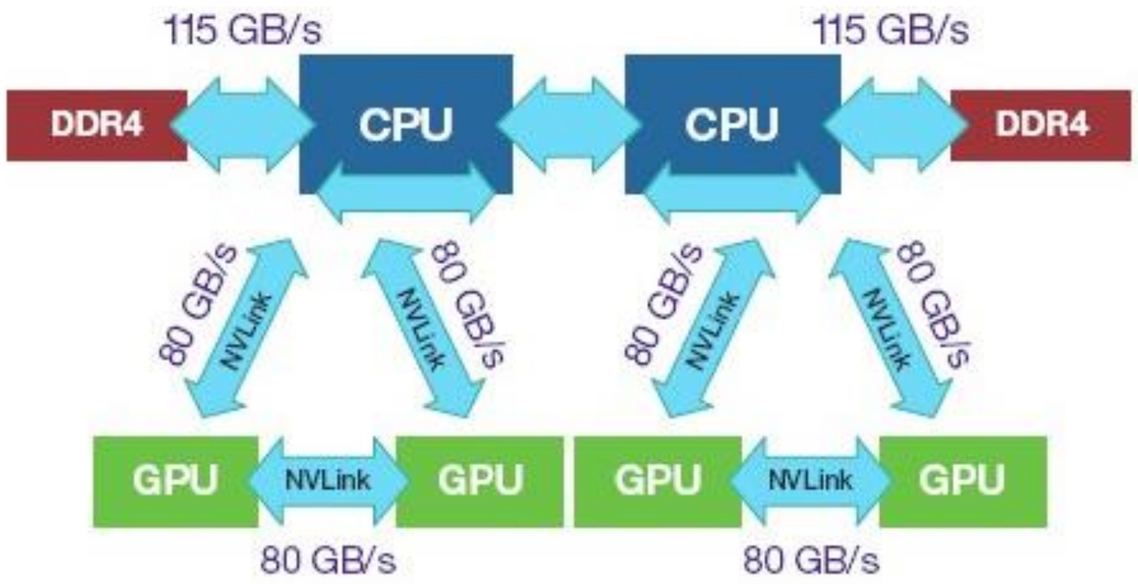
Device conductance

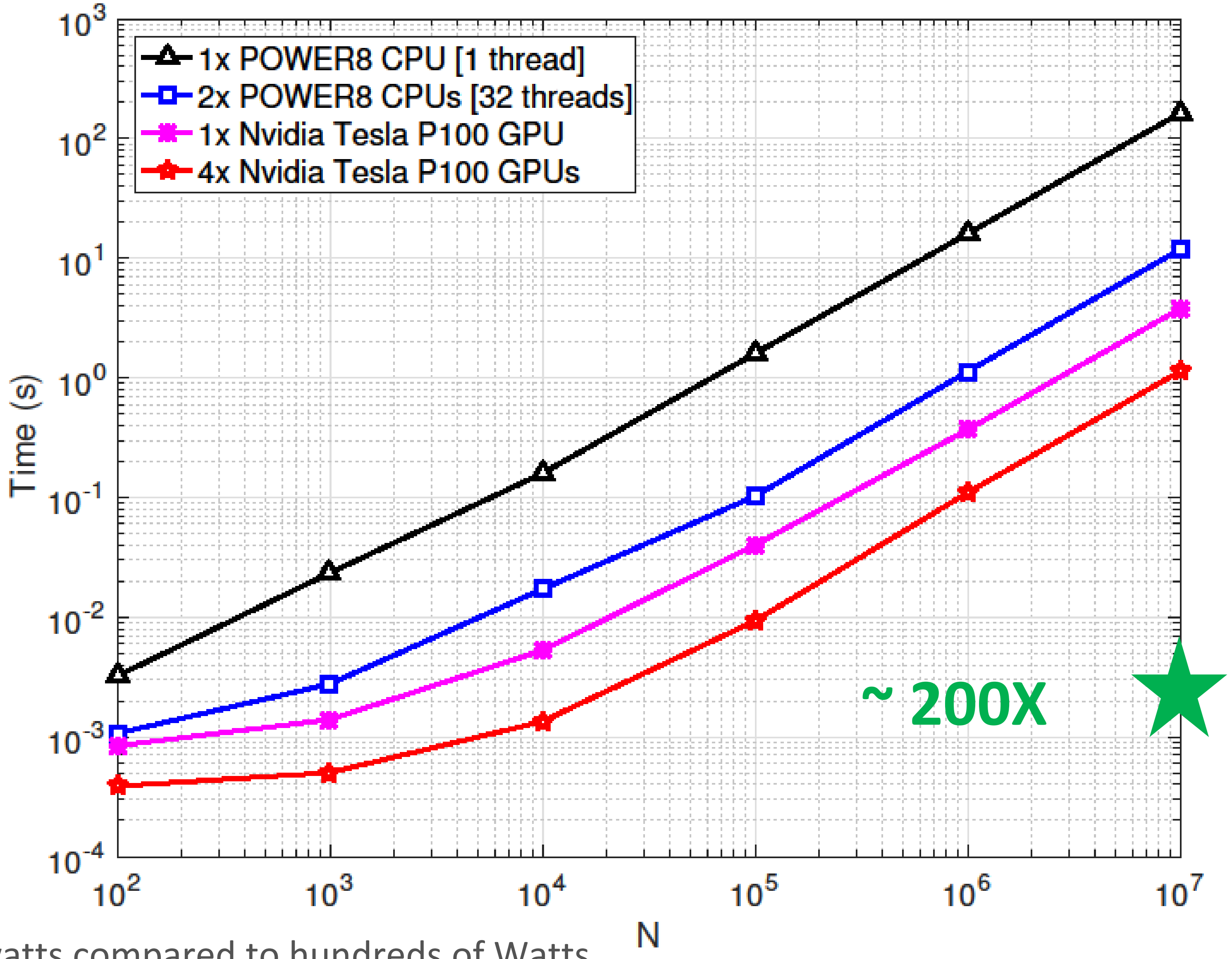
Sebastian et al., Nature Communications, 2017

Very weak correlation of c = 0.01
No shuttling back and forth of data
Massively parallel
Unprecedented areal/power efficiency

Accelerating Temporal Correlation Detection

IBM Minsky





Peak dynamic power in the order of watts compared to hundreds of Watts

Summary

DATA

Compute in Place

Move Way Faster

Store Much Denser

Thank You

1 **Visit us @ Booth #A09**

1-on-1 meetings

Personalized solution demos

2 **Explore IBM Systems solutions**

Visit: ibm.com/systems

3 **MSP Hub – discover helpful resources**

Visit: www.themsphub.com

Notices and Disclaimers

Notices and Disclaimers Con’t.

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. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM’s products. IBM EXPRESSLY DISCLAIMS ALL WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE.

The provision of the information contained h erein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, ibm.com, Aspera®, Bluemix, Blueworks Live, CICS, Clearcase, Cognos®, DOORS®, Emptoris®, Enterprise Document Management System™, FASP®, FileNet®, Global Business Services ®, Global Technology Services ®, IBM ExperienceOne™, IBM SmartCloud®, IBM Social Business®, Information on Demand, ILOG, Maximo®, MQIntegrator®, MQSeries®, Netcool®, OMEGAMON, OpenPower, PureAnalytics™, PureApplication®, pureCluster™, PureCoverage®, PureData®, PureExperience®, PureFlex®, pureQuery®, pureScale®, PureSystems®, QRadar®, Rational®, Rhapsody®, Smarter Commerce®, SoDA, SPSS, Sterling Commerce®, StoredIQ, Tealeaf®, Tivoli®, Trusteer®, Unica®, urban{code}®, Watson, WebSphere®, Worklight®, X-Force® and System z® Z/OS, are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.