WIEŚCI Z KONFERENCJI

FAST '19, VAULT '19 I MSST '19

06.06.2019, Andrzej Jackowski

USENIX FAST '19

17th USENIX Conference on File and Storage Technologies

February 25-28, Boston MA

USENIX VAULT '19

2019 Linux Storage and Filesystems Conference

February 25-26, Boston MA

MSST '19

35th International Conference on Massive Storage Systems and Technology

May 20 — 24, Santa Clara CA

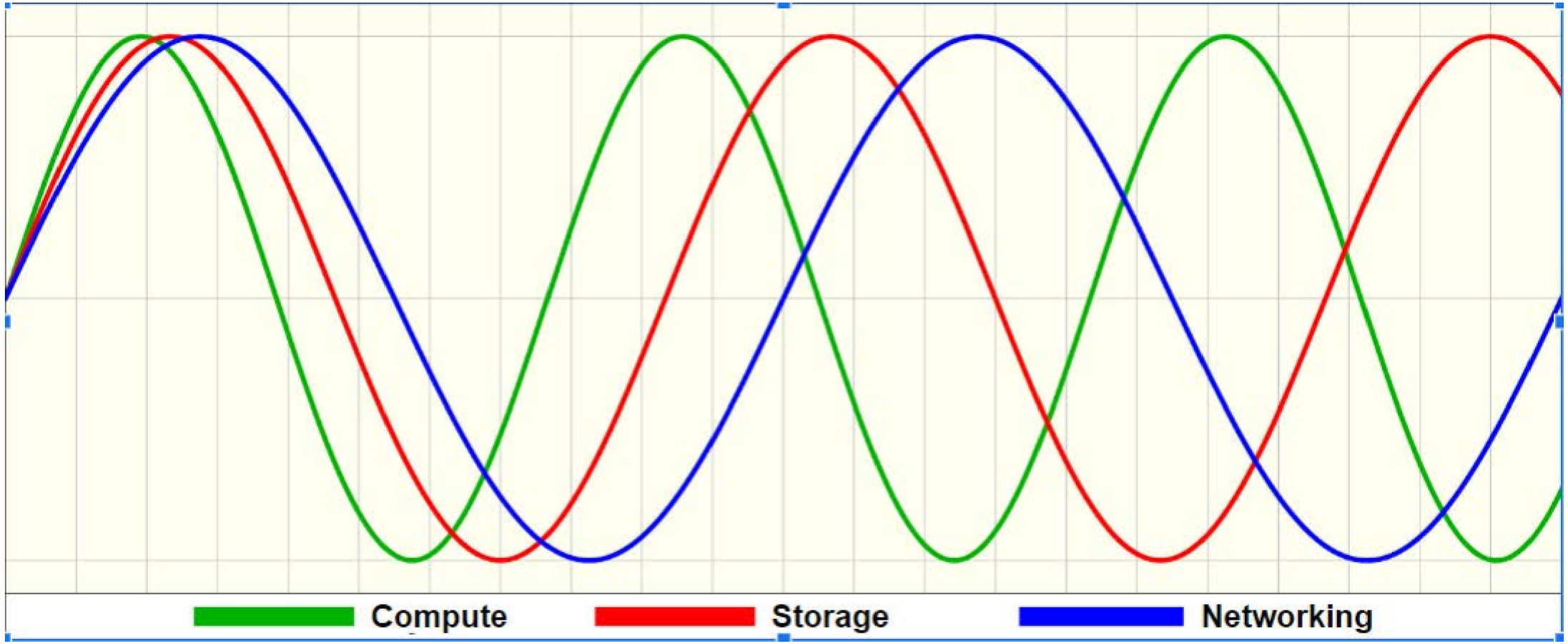
PLAN PREZENTACJI

1. Wstęp - po co nam nowe typy pamięci
2. Przegląd ciekawych publikacji
3. Sci-Fi: odległa i niepewna przyszłość

L1 cache reference	0.5	ns			
Branch mispredict	5	ns			
L2 cache reference	7	ns			14x L1 cache
Mutex lock/unlock	25	ns			
Main memory reference	100	ns			20x L2 cache, 200x L1 cache
Compress 1K bytes with Zippy	3,000	ns	3	us	
Send 1K bytes over 1 Gbps network	10,000	ns	10	us	
Read 4K randomly from SSD*	150,000	ns	150	us	~1GB/sec SSD
Read 1 MB sequentially from memory	250,000	ns	250	us	
Round trip within same datacenter	500,000	ns	500	us	
Read 1 MB sequentially from SSD*	1,000,000	ns	1,000	us	1 ms ~1GB/sec SSD, 4X memory
Disk seek	10,000,000	ns	10,000	us	10 ms 20x datacenter roundtrip
Read 1 MB sequentially from disk	20,000,000	ns	20,000	us	20 ms 80x memory, 20X SSD
Send packet CA->Netherlands->CA	150,000,000	ns	150,000	us	150 ms

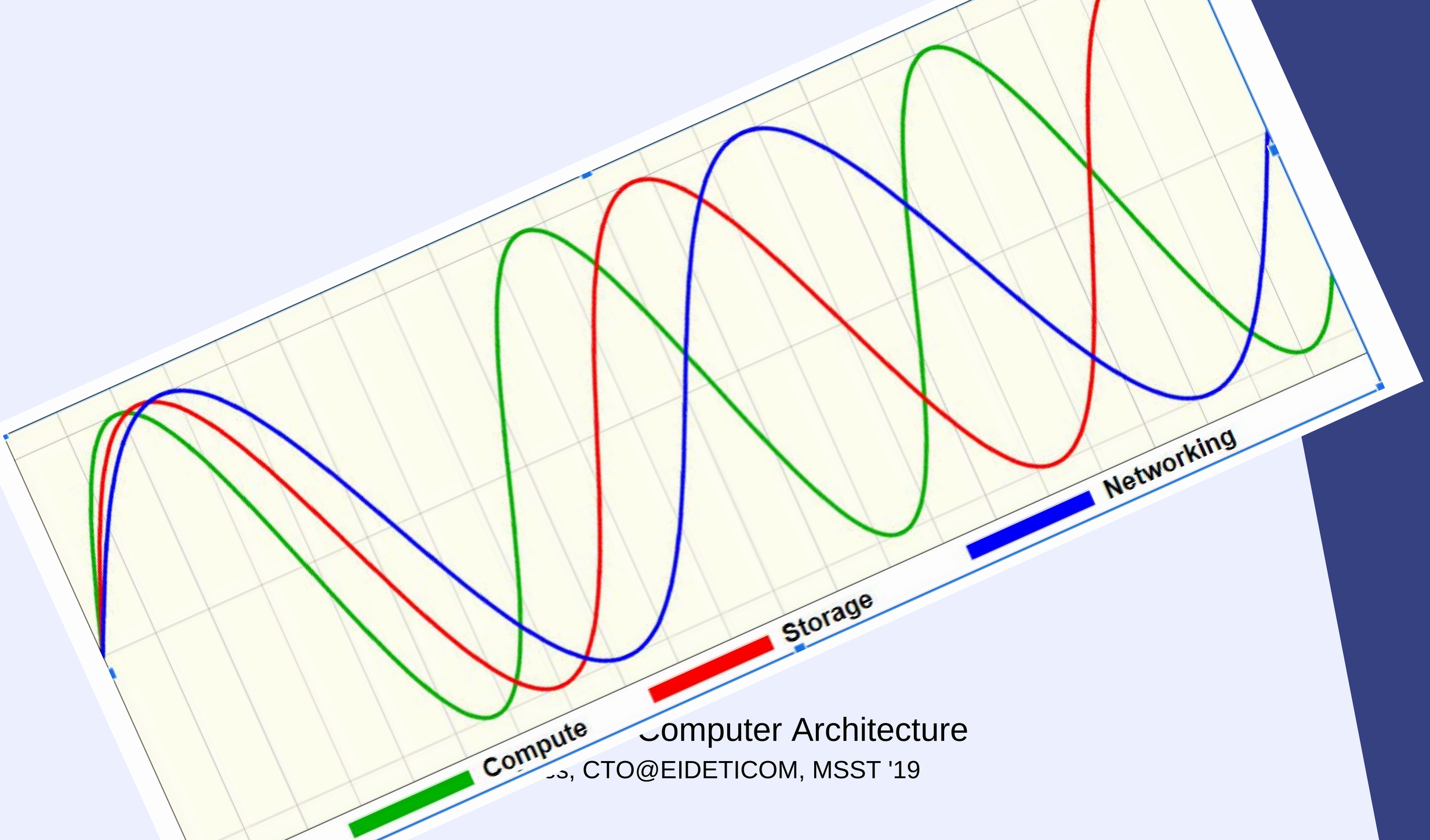
Latency Numbers Every Programmer Should Know

https://gist.github.com/jboner/2841832



The Biorhythms of Computer Architecture

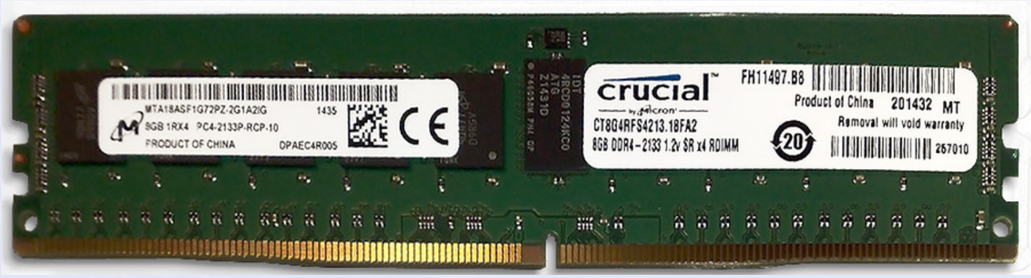
Stephen Bates, CTO@EIDETICOM, MSST '19



Computer Architecture

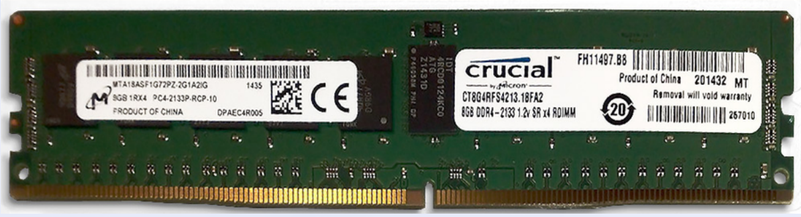
s, CTO@EIDETICOM, MSST '19

AMD Phenom II
AMD
DIFFUSED IN GERMANY
MADE IN MALAYSIA


crucial
FH11497.B8
Product of China 201432 MT
Removal will void warranty
CT8G4RFS4213.18FA2
8GB DDR4-2133 1.2v SR x4 RDIMM
MTA18ASF1G72PZ-2G1A2IG
8GB 1RX4 PC4-2133P-RCP-10
PRODUCT OF CHINA
DPAEC4R005

ADATA
SSD
2.5" SATA 6Gb/s
XPG SX900
128GB

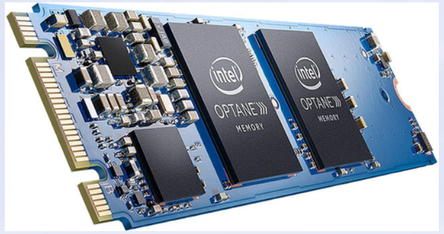
HDD
160 GB SATA
S/N: HDDR2340023948
Product of World



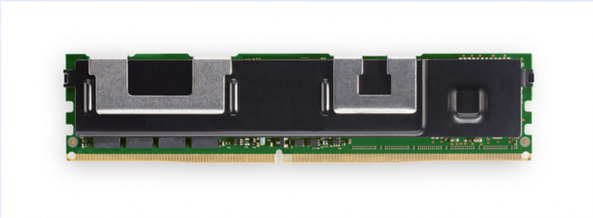
WYSOKA POJEMONOŚĆ
NIEZŁA WYDAJNOŚĆ

WYSOKA WYDAJNOŚĆ
NIEULOTNOŚĆ

M.2 SSD



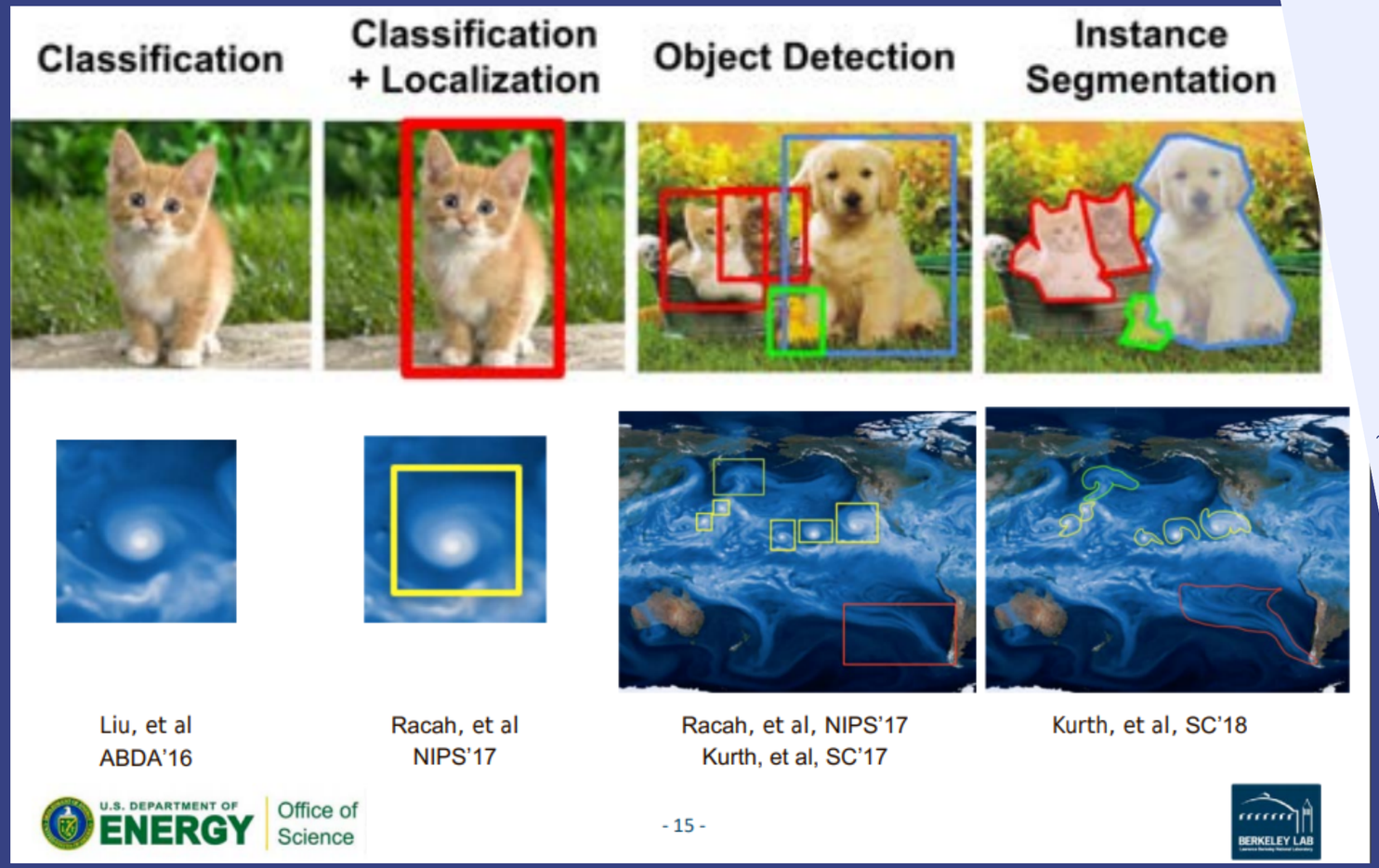
M.2 Optane



Optane DIMM

NVDIMM-N

National Energy Research Scientific Computing Center



27,360 VOLTA GPUs

1.13EF peak performance (16-bit)

~61 MB/sample × ~65,000 samples/s @ 27K GPUs → ~3.8 TB/s

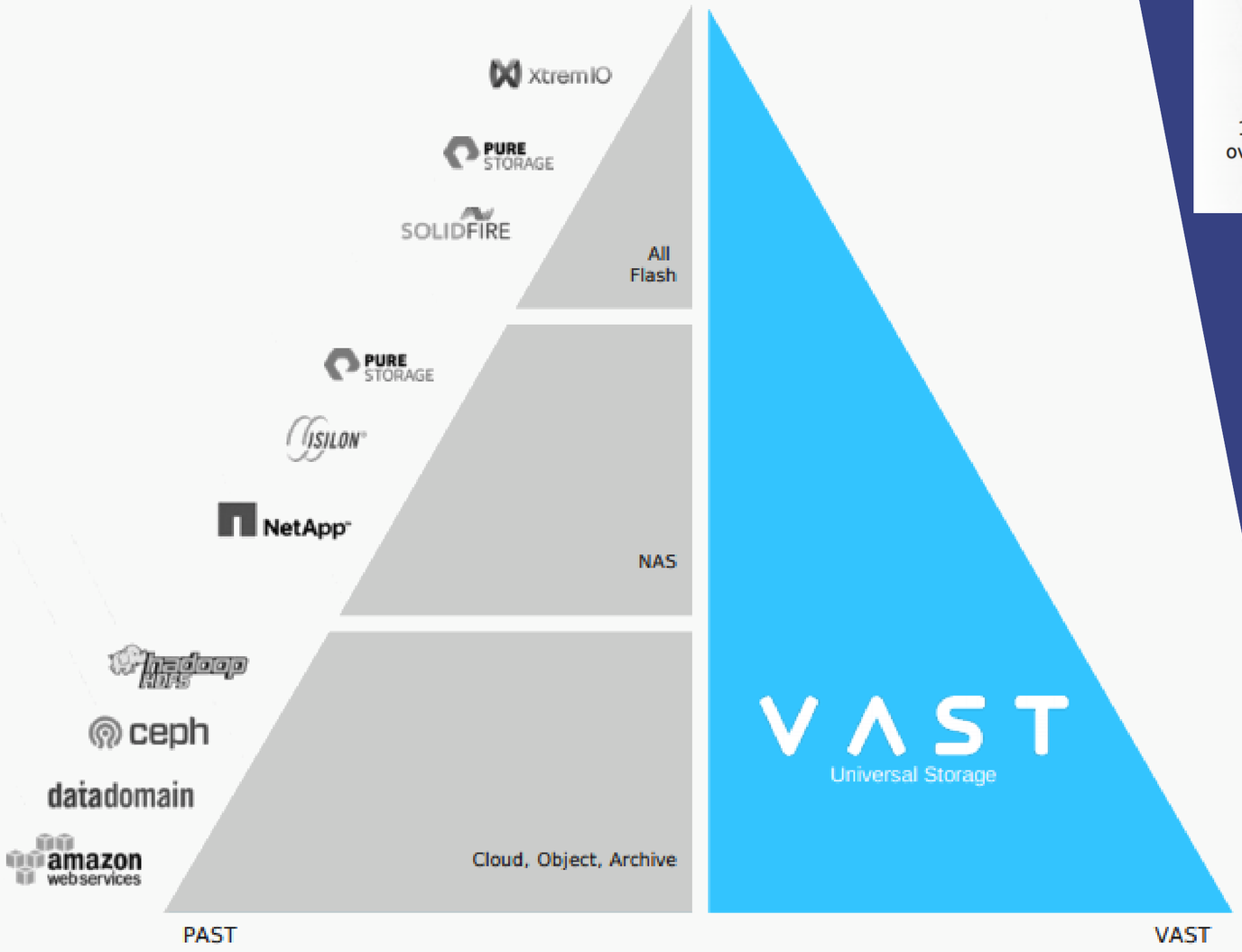
– Typical distributed FS bandwidth: ~400 GB/s

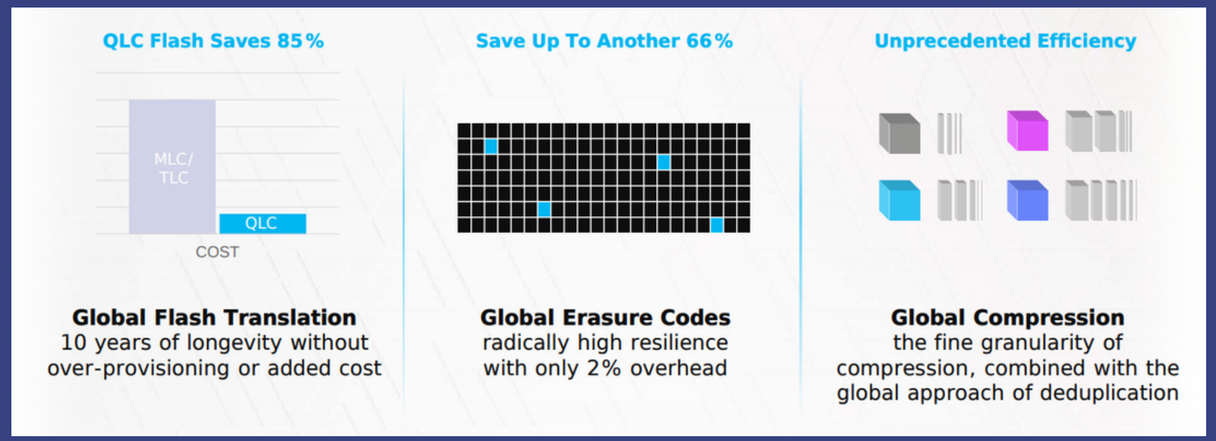
~8x performance gap

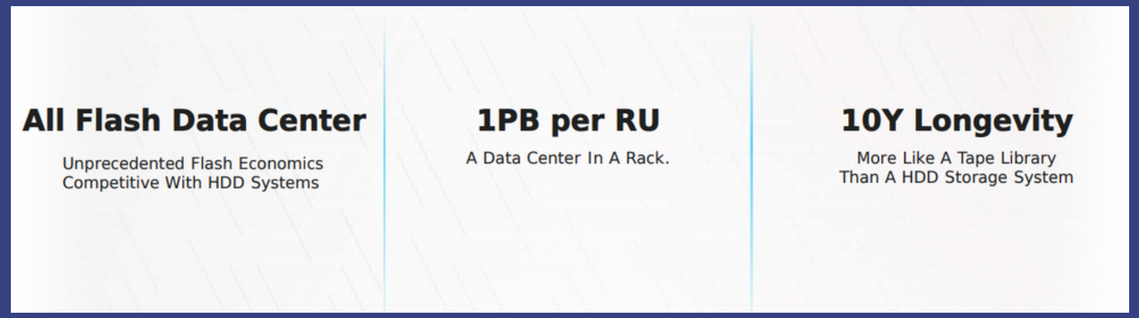
– Typical Burst Buffer bandwidth: ~2 TB/s

~2x performance gap

http://storageconference.us/2019/Invited/Koziol.slides.pdf







http://storageconference.us/2019/Invited/Hallak.slides.pdf

Część II

CIEKAWE PUBLIKACJE

DistCache: Provable Load Balancing for Large-Scale Storage Systems with Distributed Caching

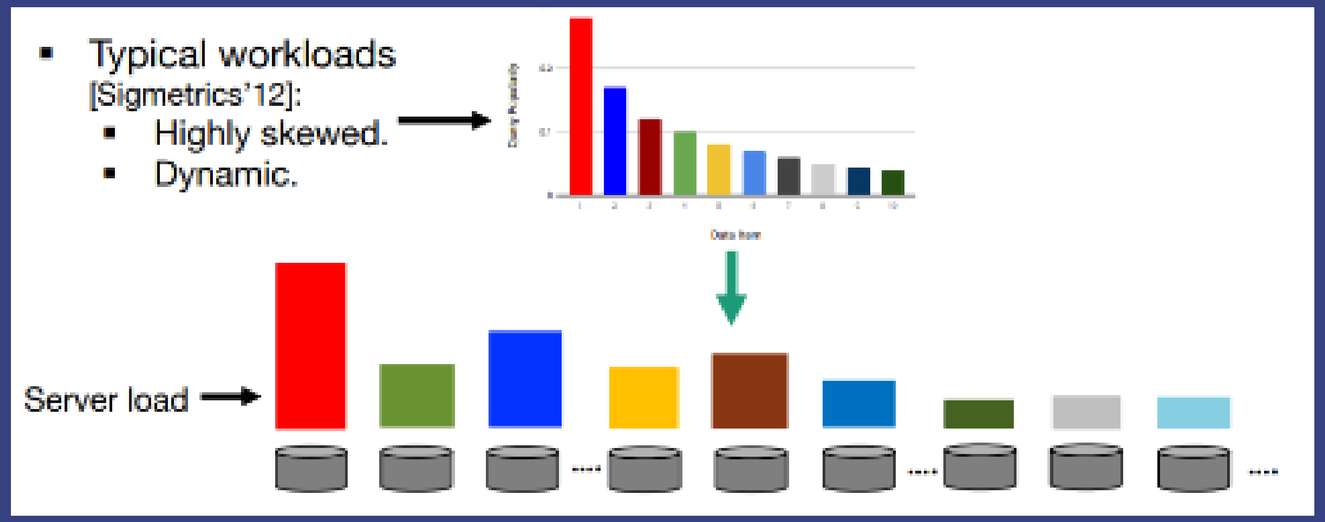
Fast '19

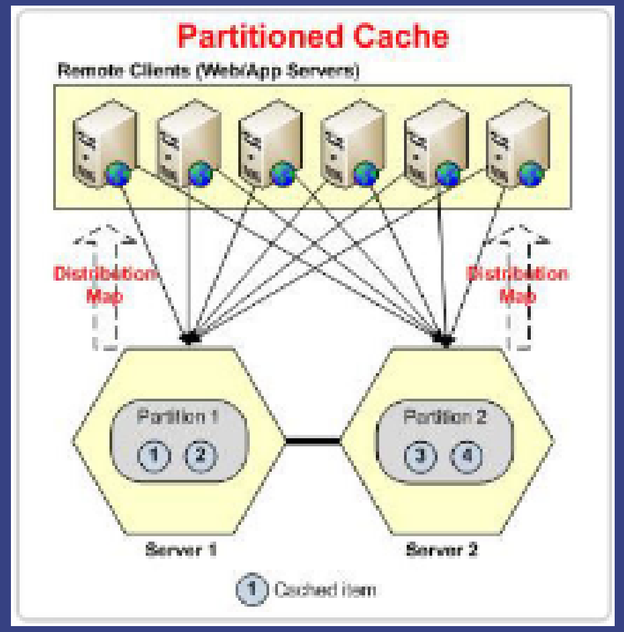
- Systemy są niezbalansowane

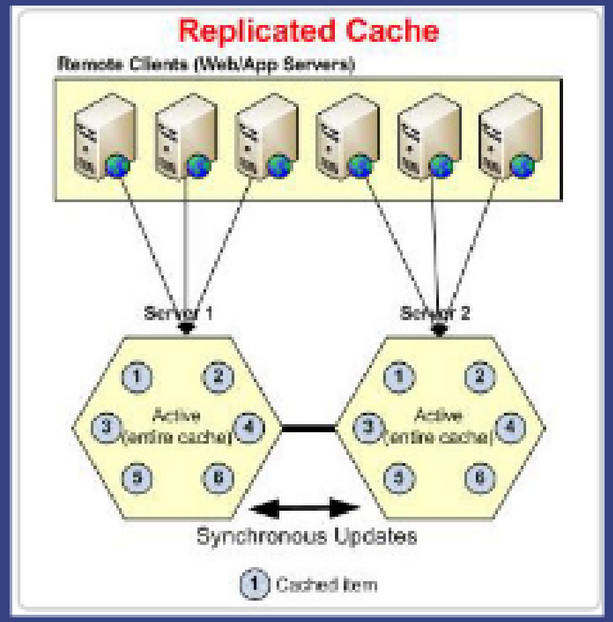
- Jeden wielki cache to problem

- Partycjonowanie (*cache partition*): wąskie gardło na jednym węźle

- Replikacja (*cache replication*): gigantyczny narzut na spójność







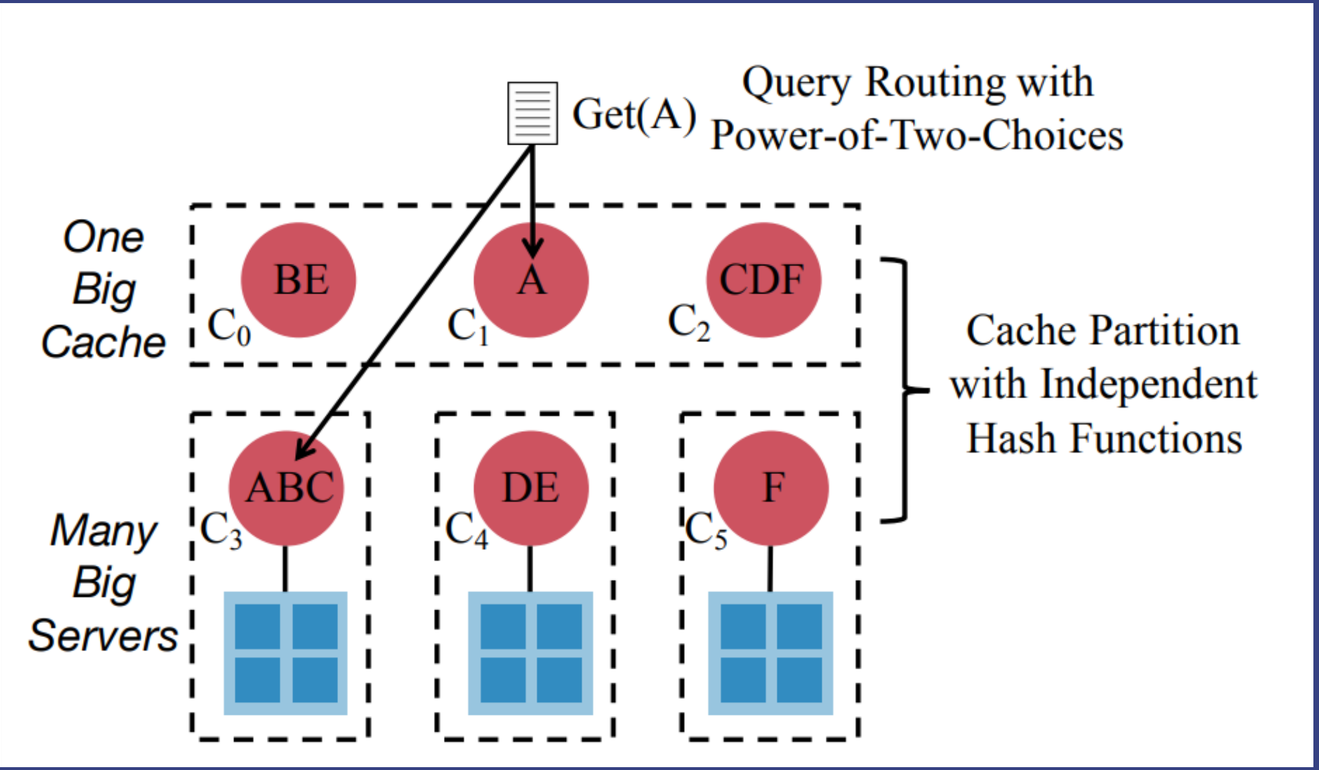
DistCache: Provable Load Balancing for Large-Scale Storage Systems with Distributed Caching

Fast '19

- Dwie warstwy cache, każda ma niezależną funkcję hashującą

- *Power-of-two-choices* między dwiema warstwami

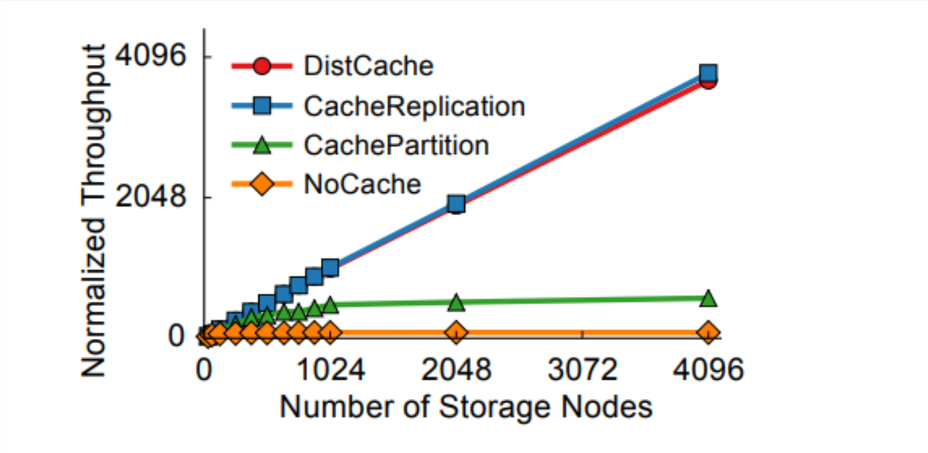
- Detale trochę bardziej skomplikowane, m.in. użyto skojarzeń w grafie dwudzielnym



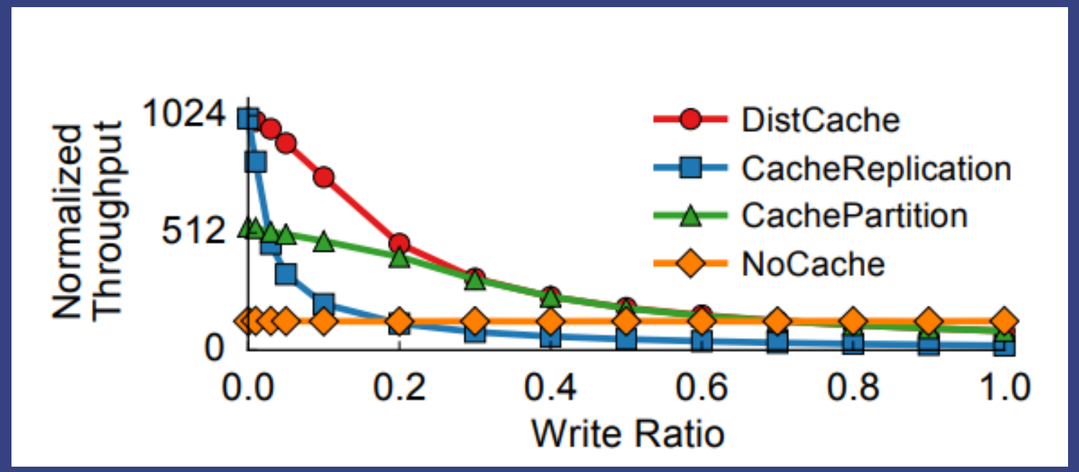
DistCache: Provable Load Balancing for Large-Scale Storage Systems with Distributed Caching

Fast '19

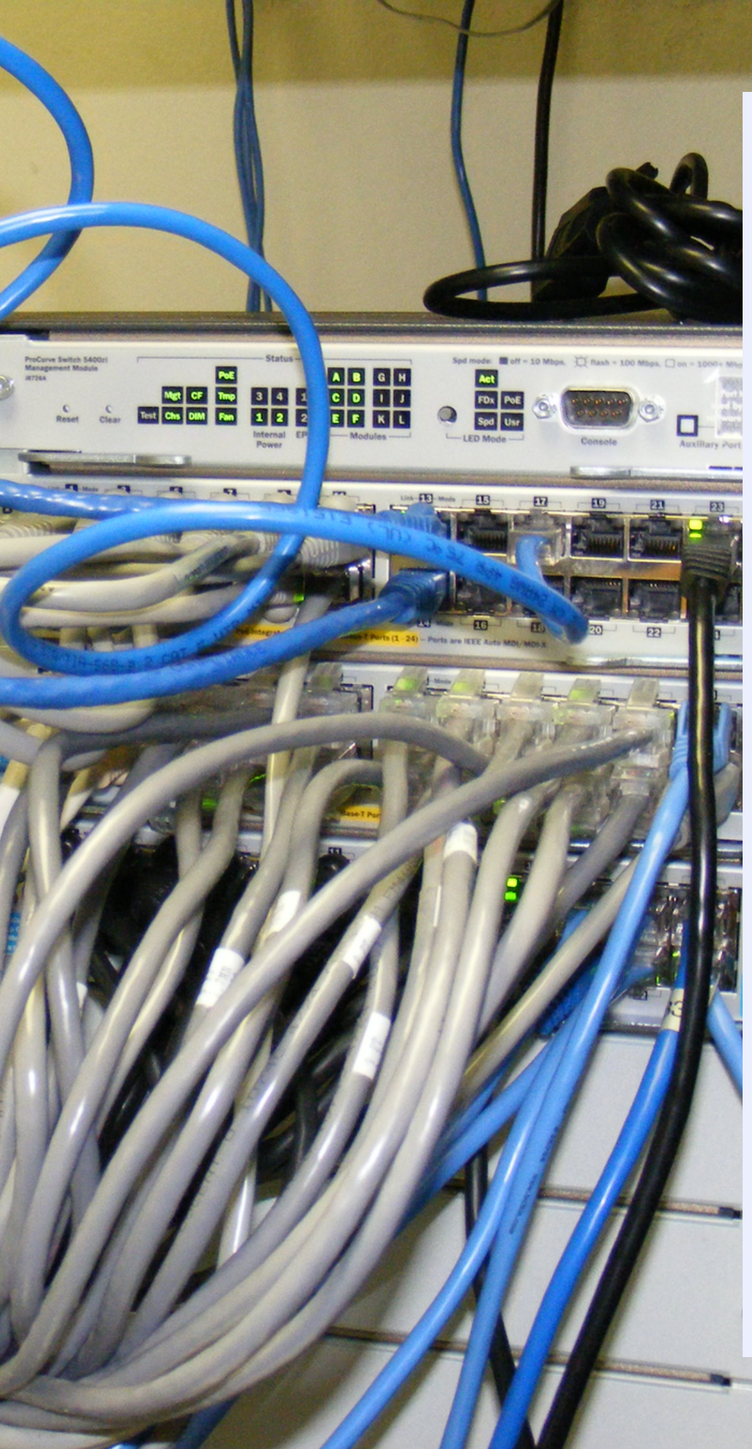
Skaluje się jak replikacja



Niski narzut jak przy partycjonowaniu



+ trochę gwarancji teoretycznych



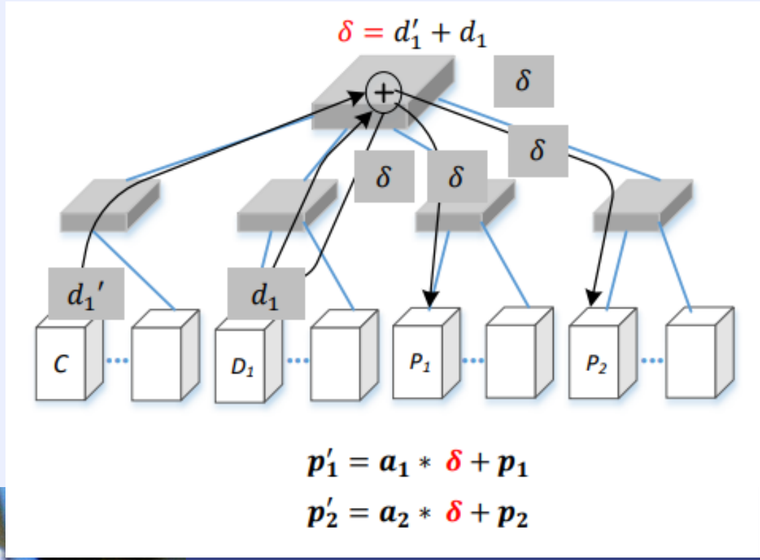
XORINC: OPTIMIZING DATA REPAIR AND UPDATE FOR ERASURE-CODED SYSTEMS WITH XOR-BASED IN-NETWORK COMPUTATION

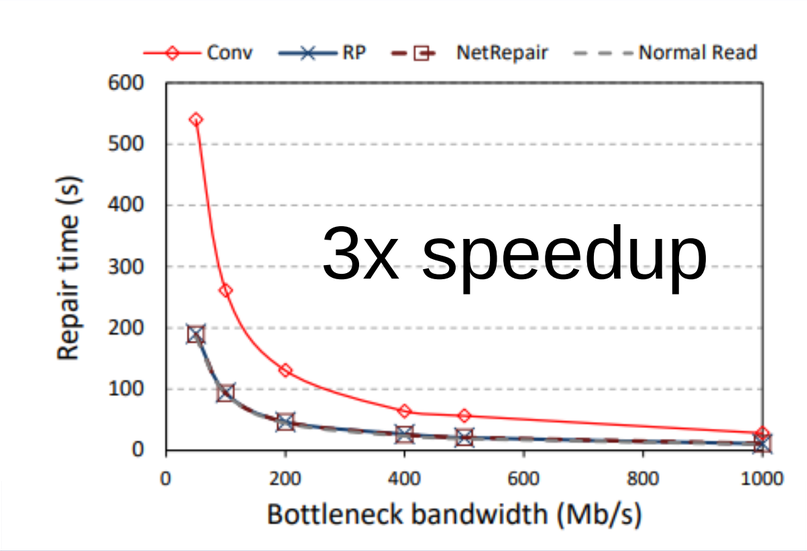
MSST '19

PROGRAMOWALNE SIECI

Pozwalają przesunąć część obliczeń do switcha, co pozwala zmniejszyć ruch sieciowy

POMYSŁ - ZASTOSOWANIE W ERASURE CODACH





ZIGGURAT: A TIERED FILE SYSTEM FOR NON-VOLATILE MAIN MEMORIES AND DISKS

FAST '19

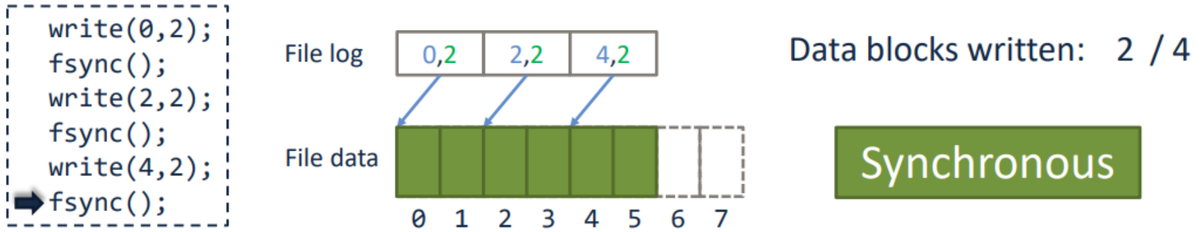
CZASAMI SSD SĄ ZA WOLNE!

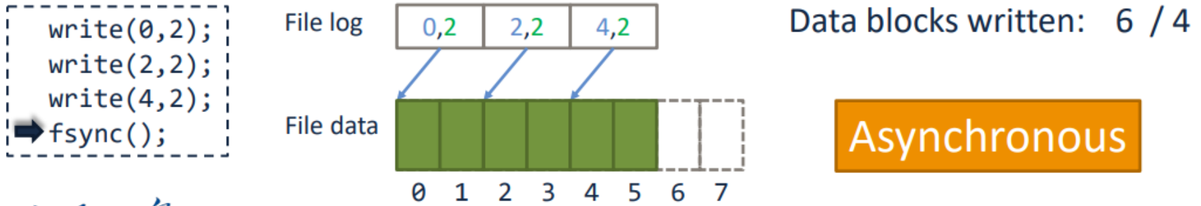
Dlatego zaprpononowano tiering z użyciem NVMM

ROZWIĄZANIE

- Wrzucanie asynchronicznych zapisów od razu na niższą warstwę
- Przerzucanie tylko zimnych plików na dyski

Data Placement		Synchronicity predictor	
		Synchronously-updated	Asynchronously-updated
Write size predictor	Large writes	NVMM	Disk
	Small writes	NVMM	NVMM

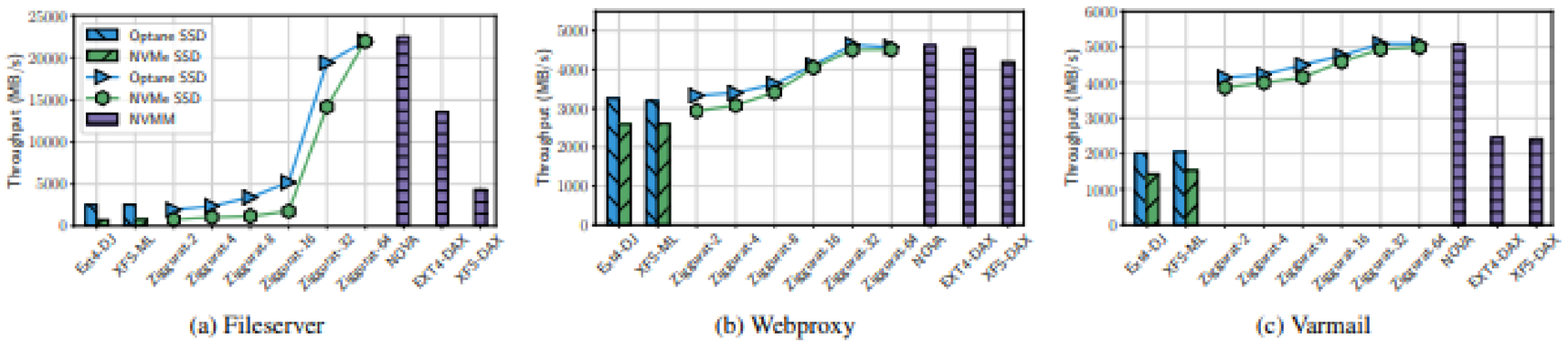




ZIGGURAT: A TIERED FILE SYSTEM FOR NON-VOLATILE MAIN MEMORIES AND DISKS

FAST '19

- 40X SZYBSZY NIŻ SAMO SSD
- DOBRE WYNIKI W TYPOWYCH ZASTOSOWANIACH (SQLITE, ROCKSDB)
- RÓŻNE WYNIKI W ZALEŻNOŚCI OD ROZMIARU PAMIĘCI NVMM (GB)
- PAMIĘĆ NVMM BYŁA EMULOWANA

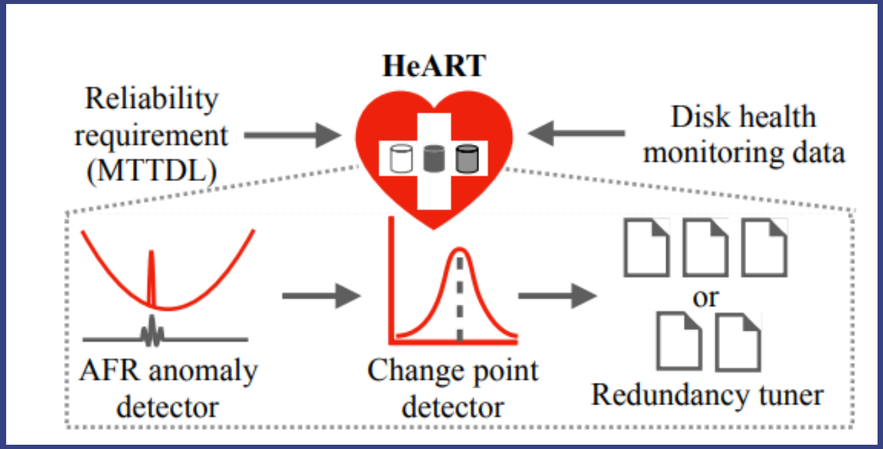


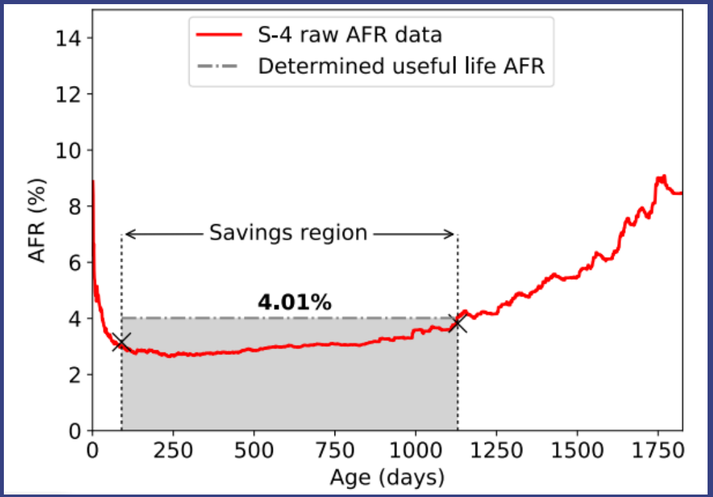
(a) Fileserver (b) Webproxy (c) Varmail

CLUSTER STORAGE SYSTEMS GOTTA HAVE HEART: IMPROVING STORAGE EFFICIENCY BY EXPLOITING DISK-RELIABILITY HETEROGENEITY
FAST '19

Dyski twarde psują się najbardziej na początku i po kilku latach

- Można ograniczyć liczbę redundantnych danych na czas gdy dyski się mniej psują
- **~10% oszczędności**
- Testowane na zbiorze danych Backblaze (100k HDDs, 5 lat)





Long-Term JPEG Data Protection and Recovery for NAND Flash-Based Solid-State Storage

MSST '19

Cztery usprawnienia:

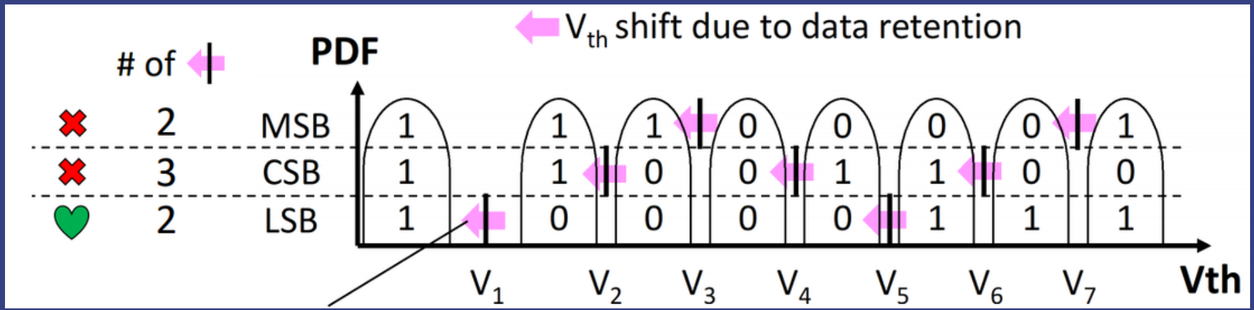
- Nagłówki przechowywane w mocniejszych bitach /stronach

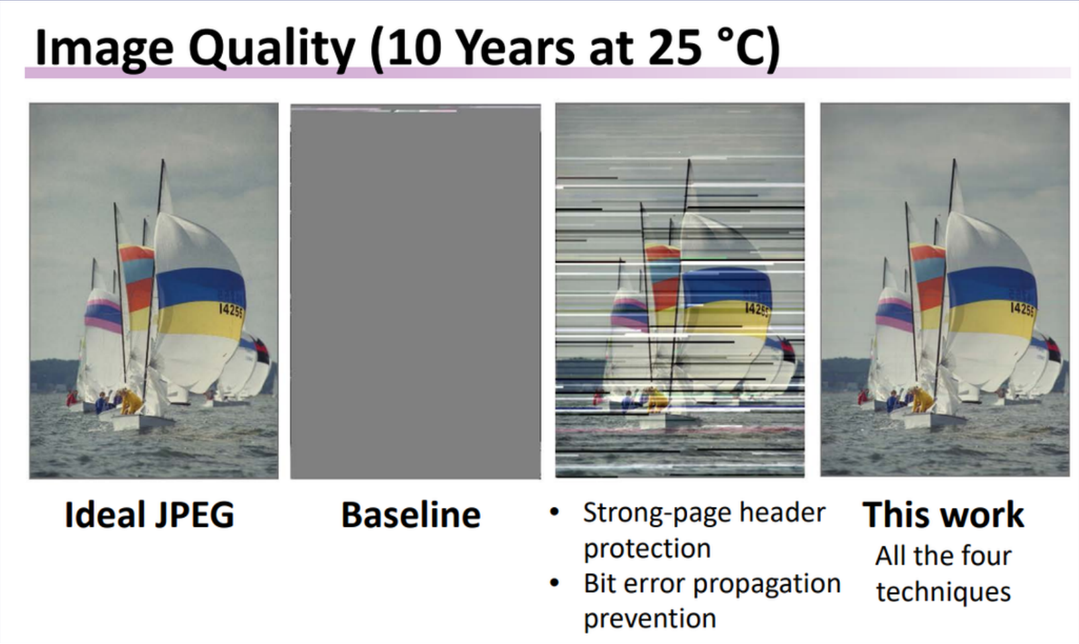
- Zapamiętywana jest długość bloków 8x8 by zapobiec propagacji

- Wykorzystanie miniaturki do korekcji błędów

- Zastosowanie kodów Huffmana

Zwiększenie rozmiaru tylko o 9.9%

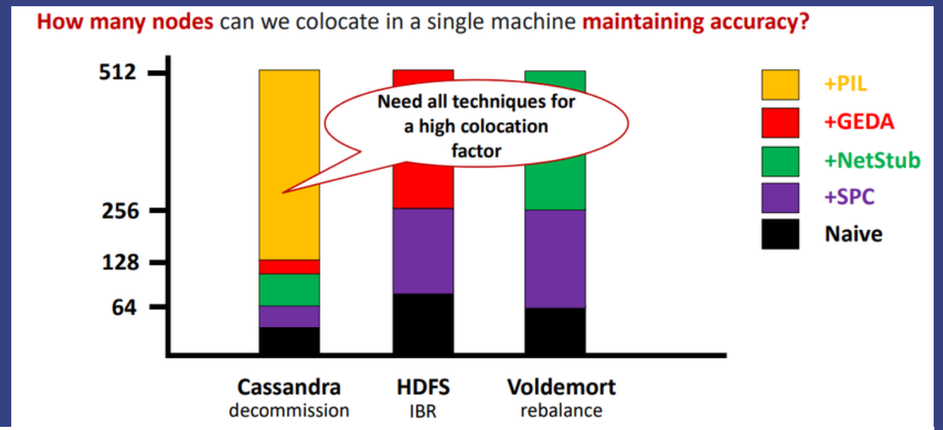


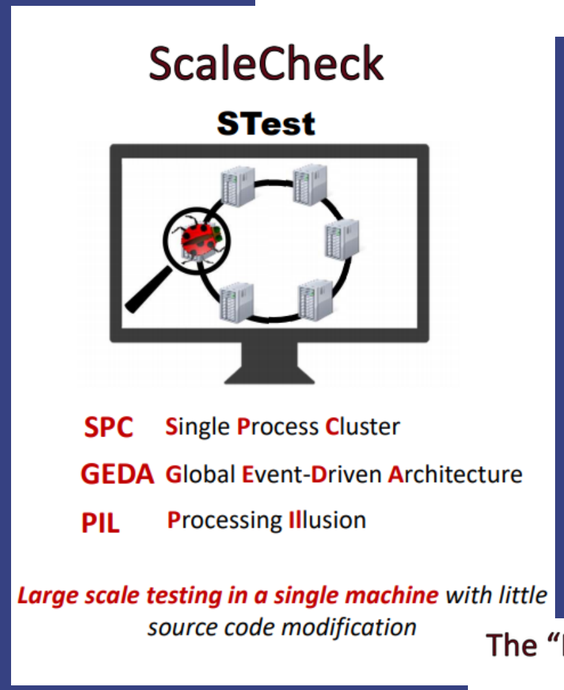


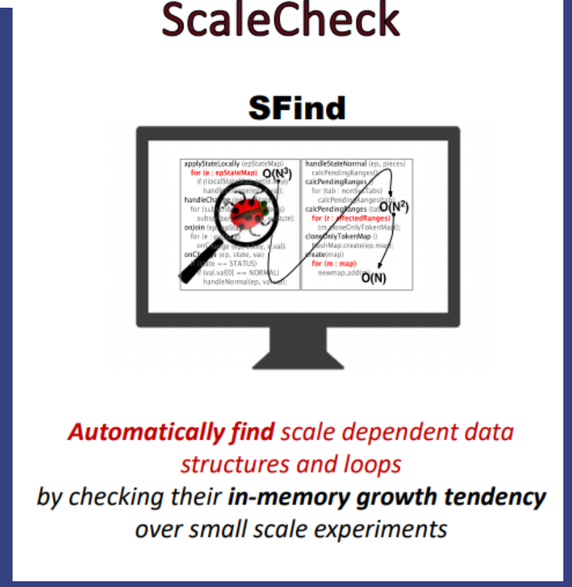
ScaleCheck: A Single-Machine Approach for Discovering Scalability Bugs in Large Distributed Systems
FAST '19

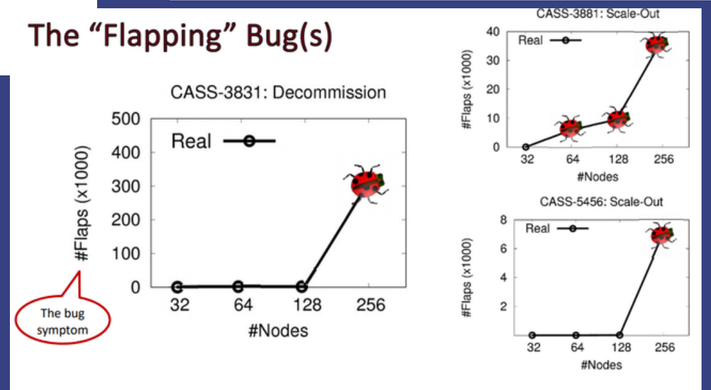
Nowy rodzaj błędów - *scalability bugs*
- Poprawki i regresja iie mogą być testowane w małym środowisku
- Problem szczególnie dla opensource
- **SFIND**: poszukuje szybko rosnących pętli i struktur danych (np. O(n^3))
- **STEST –** ułatwia testowanie całego klastra na pojedyńczej maszynie

Zreprodukowali 10 błędów i znaleźli 4 nowe (Cassandra, HDFS i więcej)

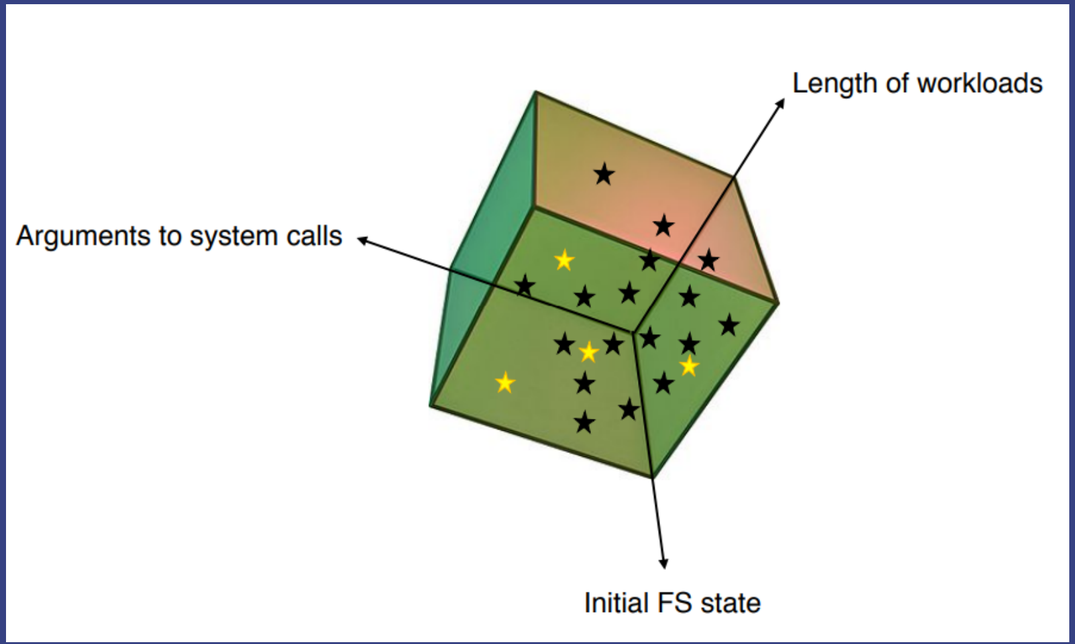








CrashMonkey: Finding File System Crash-Consistency Bugs with Bounded Black-Box Testing
Vault '19



Framework do testowania spójności systemu plików podczas crashów

Automatycznie generuje testy dla danych ograniczeń

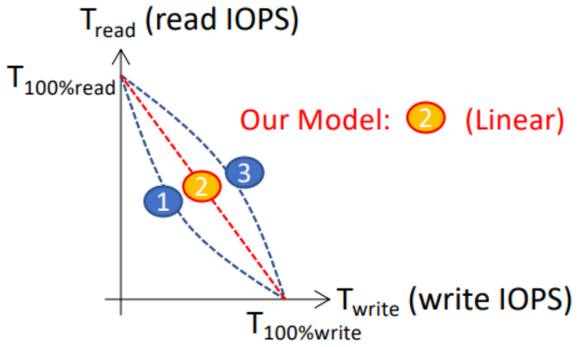
10 nowych błędów w btrfs i F2FS
Znaleźli błąd w **formalnie zweryfikowanym** FSCQ

Dostępny na GitHub:
https://github.com/utsaslab/crashmonkey

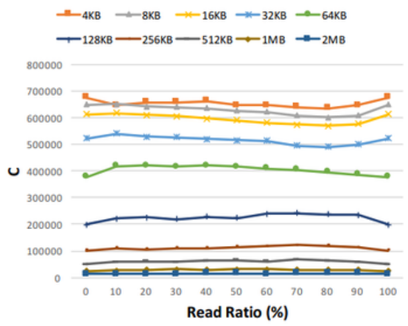
FIGHTING WITH UNKNOWNS: ESTIMATING THE PERFORMANCE OF SCALABLE DISTRIBUTED STORAGE SYSTEMS WITH MINIMAL MEASUREMENT DATA

MSST '19

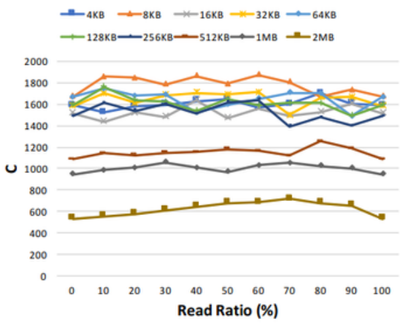
$$f_{rw} = \frac{T_{100\%read}}{T_{100\%write}}$$



$$C = T_{read} + T_{write} \cdot f_{rw}$$

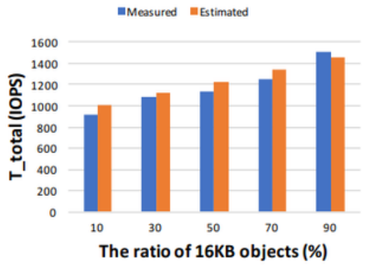


(a) Ceph

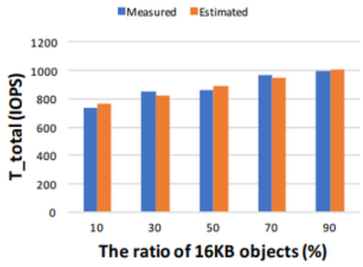


(b) Swift

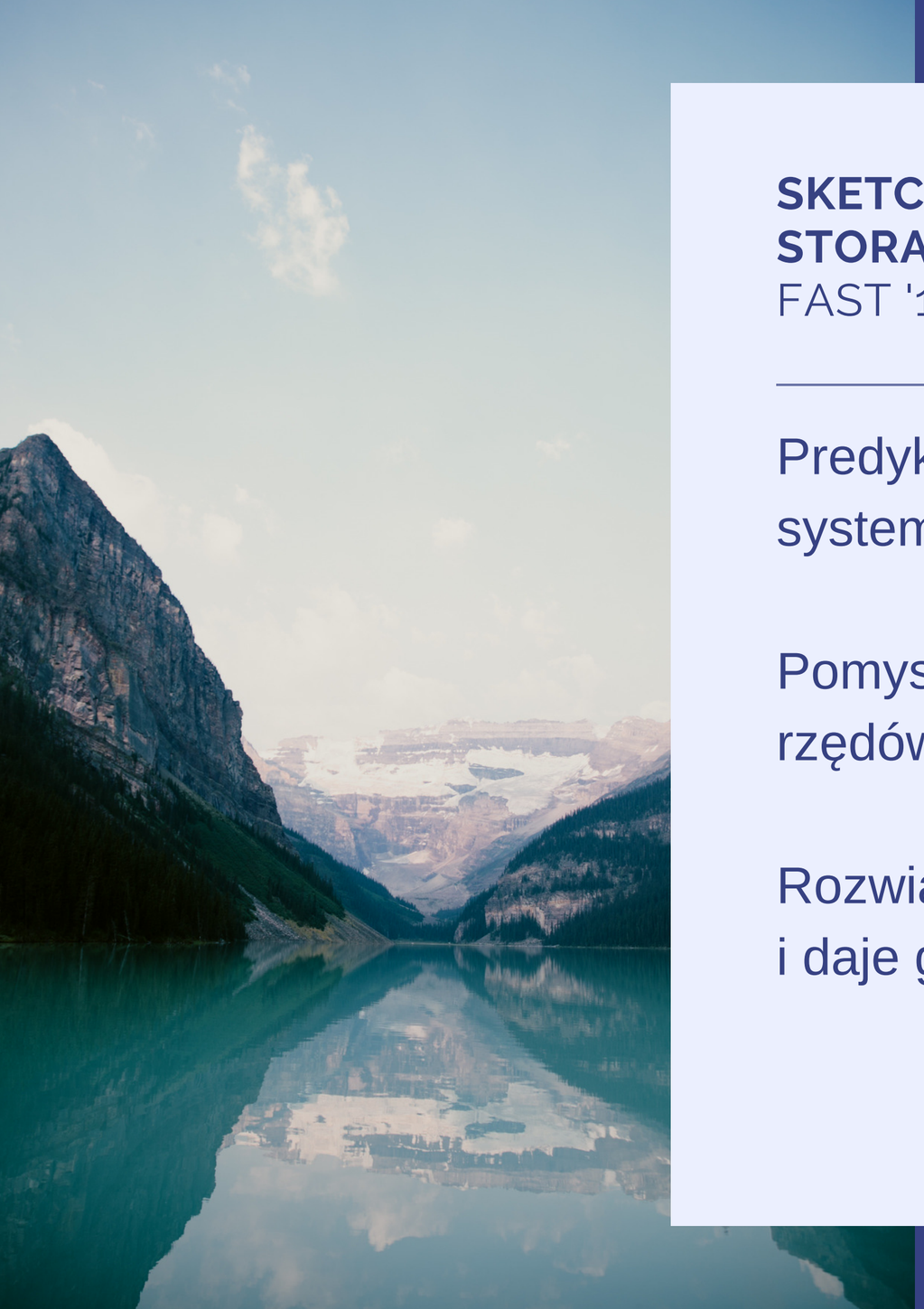
i=1



(a) 16KB Read+1MB Read



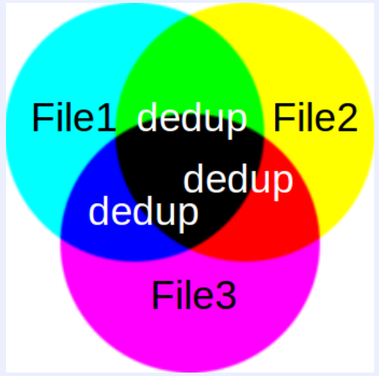
(b) 16KB RW+512KB RW



SKETCHING VOLUME CAPACITIES IN DEDUPLICATED STORAGE

FAST '19

Predykcja zwolnionego miejsca w systemach z deduplikacją jest trudna.



Pomysł - zmniejszenie problemu o wiele rzędów wielkości przy pomocy *sketchy*

Rozwiązanie działa w praktyce i daje gwarancje teoretyczne

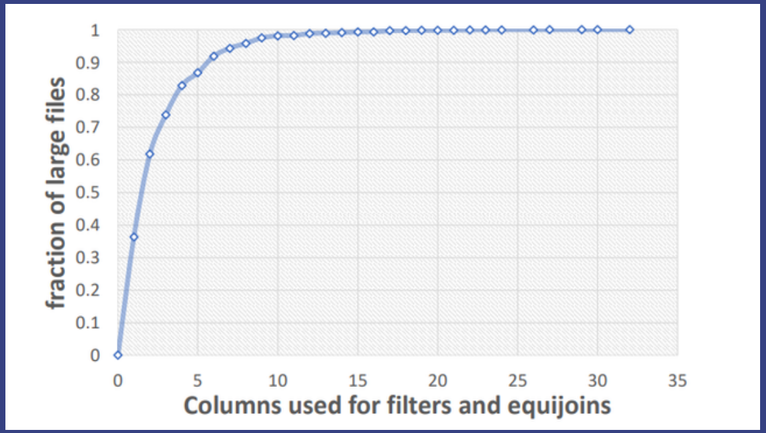
INSTalytics: Cluster Filesystem Co-design for Big-data Analytics

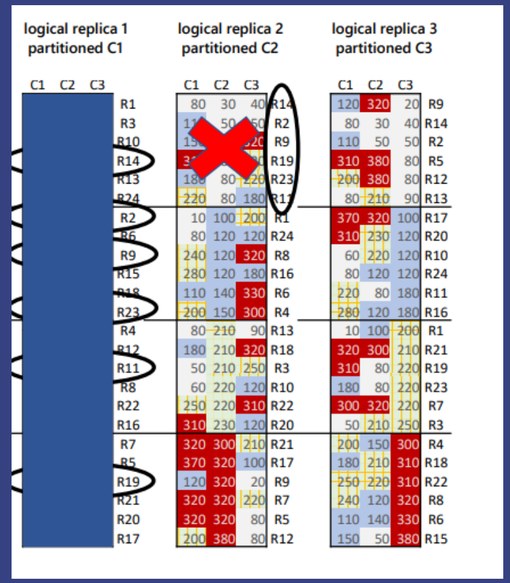
FAST '19

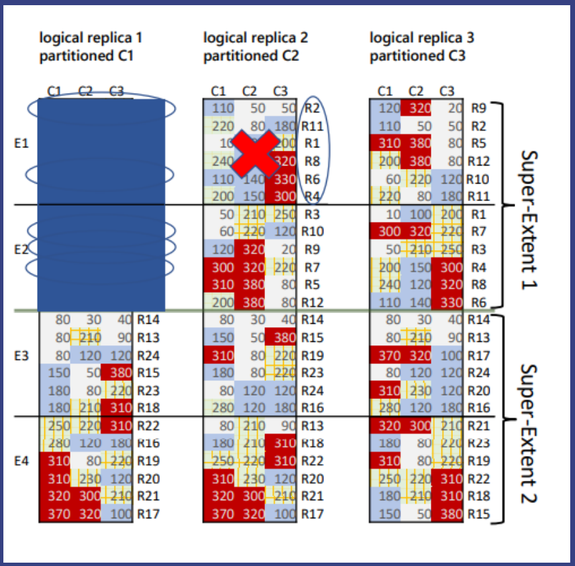
Przyspieszenie Big Data - sortowanie po różnych kluczach w replikach

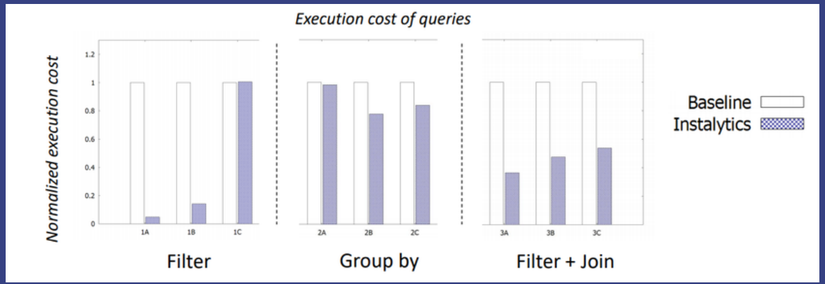
Naiwne rozwiązanie powodowało zbyt dużo rekonstrukcji

Układając sprytnie dane można naprawić rekonstrukcje i przyspieszyć wykonywanie zapytań bez dodatkowych kosztów







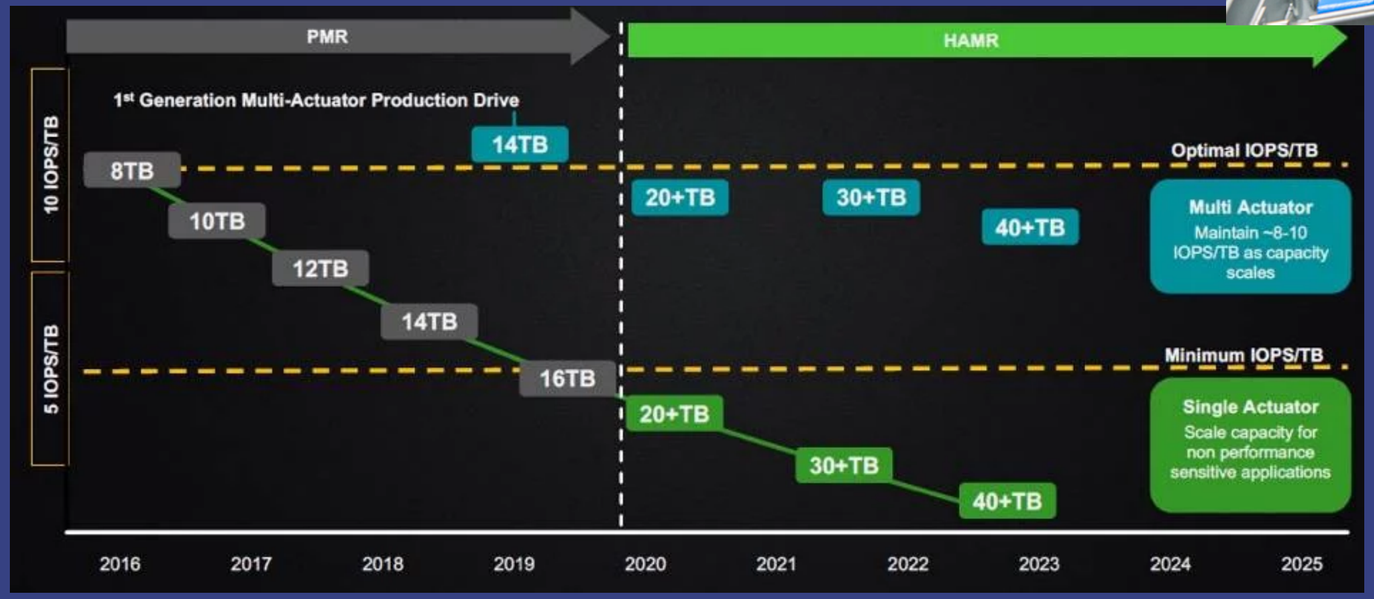


Część III

SCI-FI

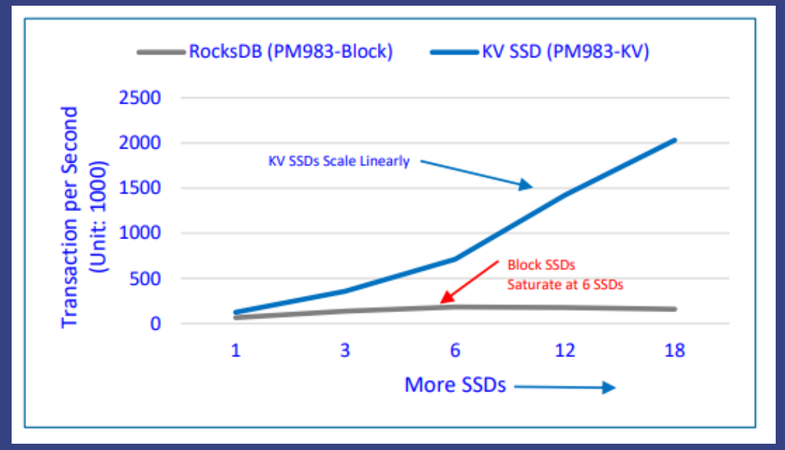
A Perspective on the Past and Future of Magnetic Hard Drives

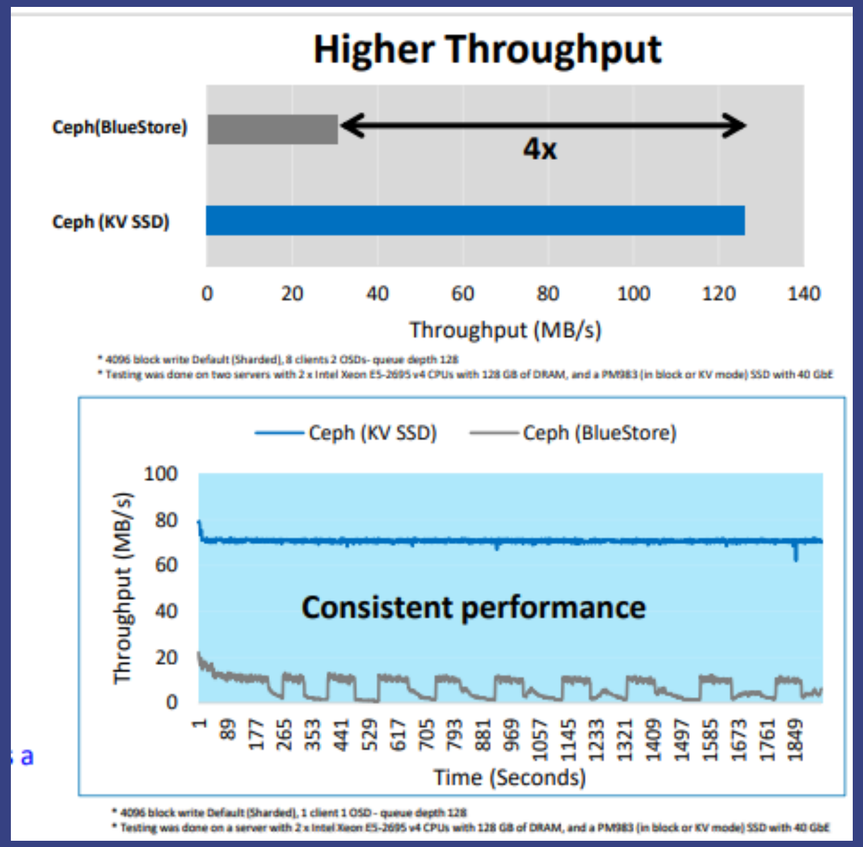
MSST '19

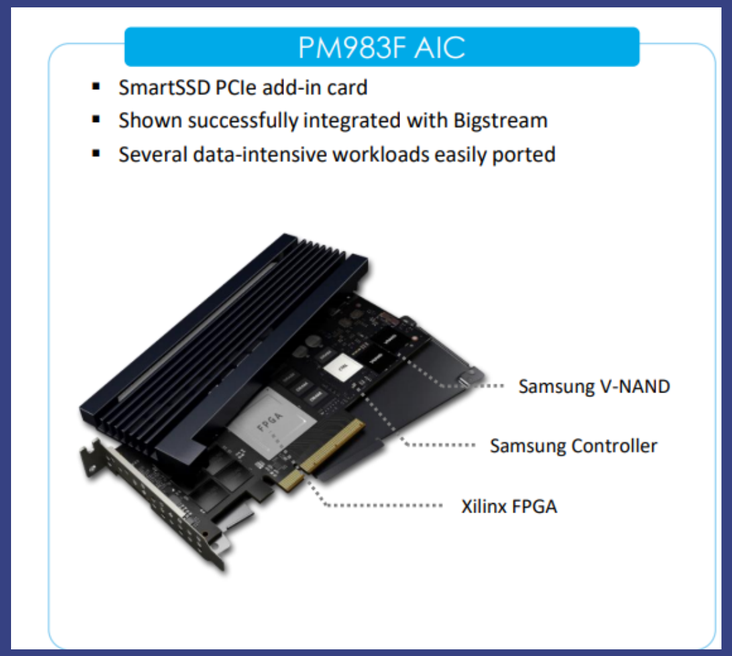


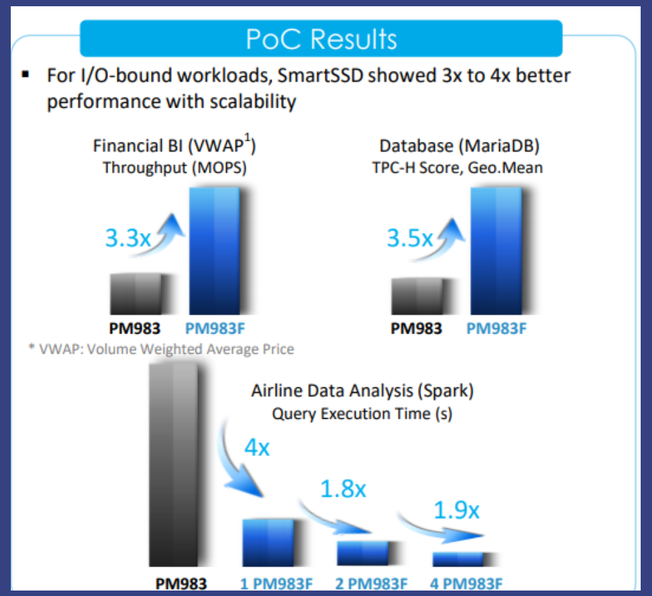
Storage in the New Age of AI/ML

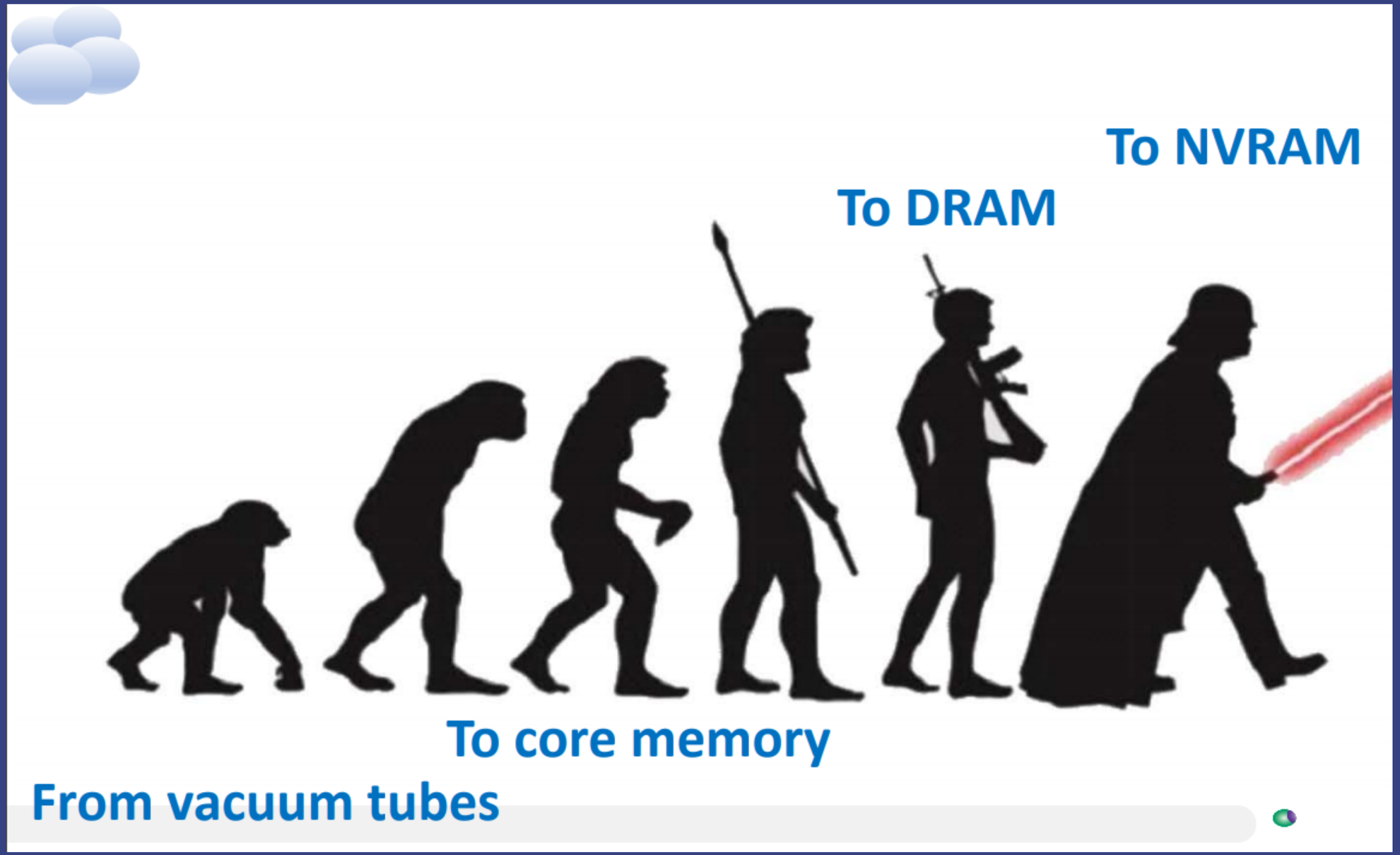
MSST '19

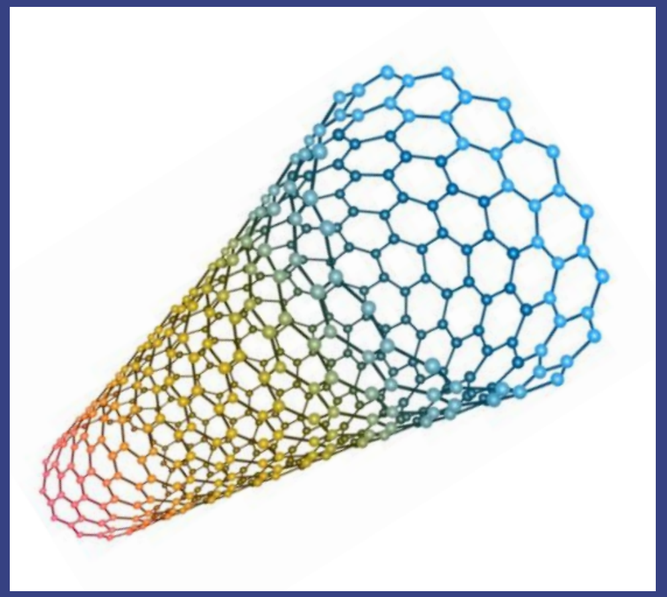






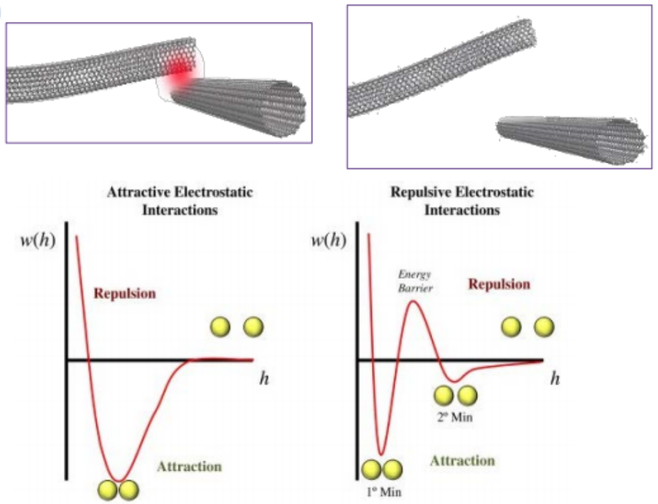


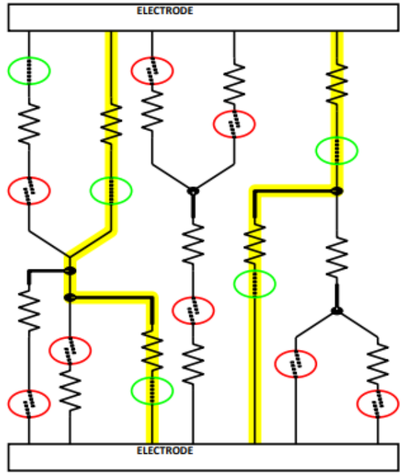

To NVRAM
To DRAM
To core memory
From vacuum tubes



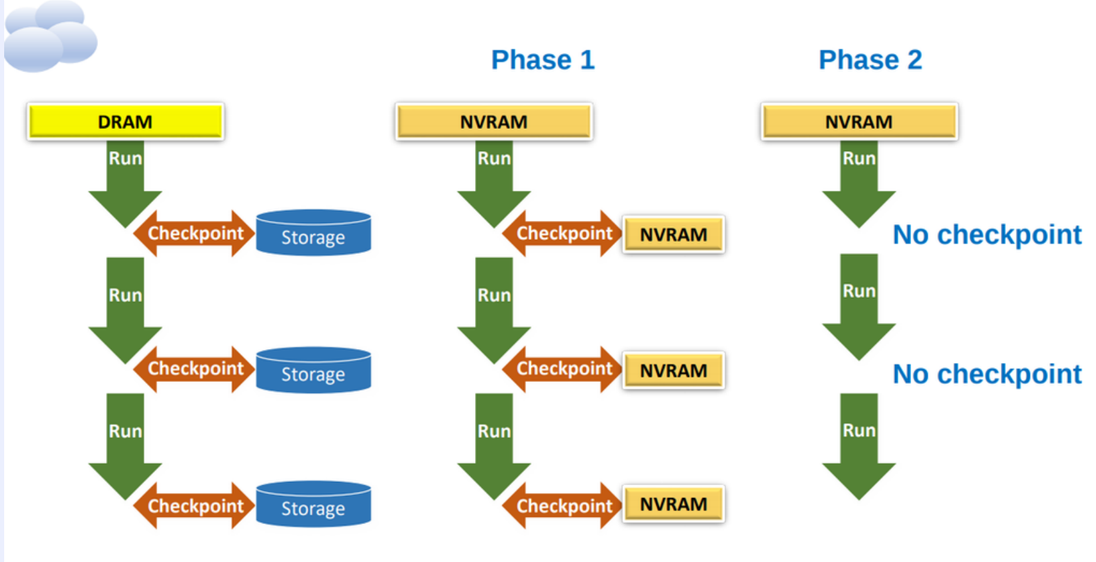
NVRAM

- DRAM speed
- Non-volatility
- Unlimited write endurance
- Wide temperature range
- Scalable beyond DRAM
- Flexible fabrication & application
- Low power
- Low cost

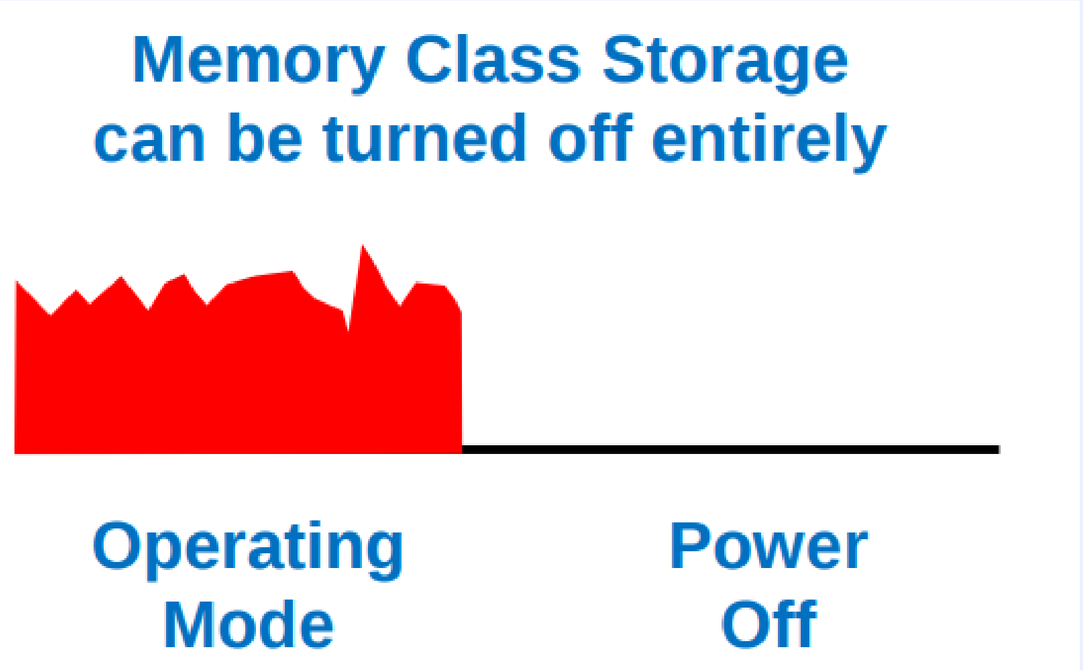


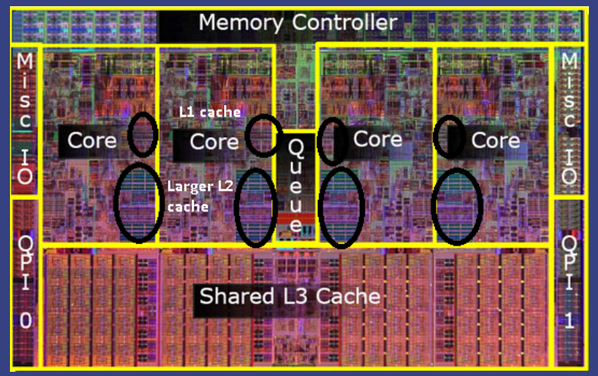


NANTERO


Phase 1
Phase 2
DRAM
NVRAM
NVRAM
Run
Checkpoint
Storage
NVRAM
No checkpoint
No checkpoint

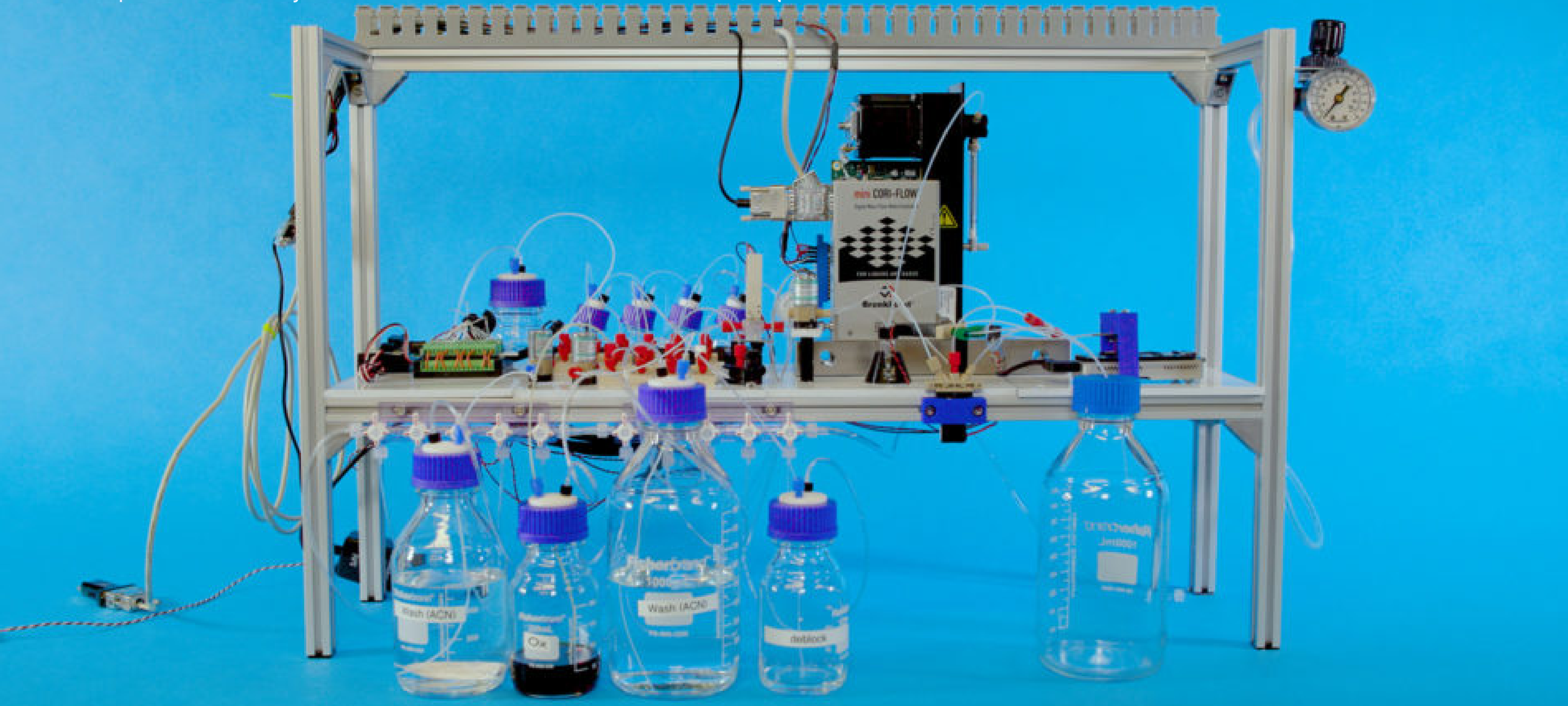
Restarting


Memory Class Storage
can be turned off entirely
Operating
Mode
Power
Off

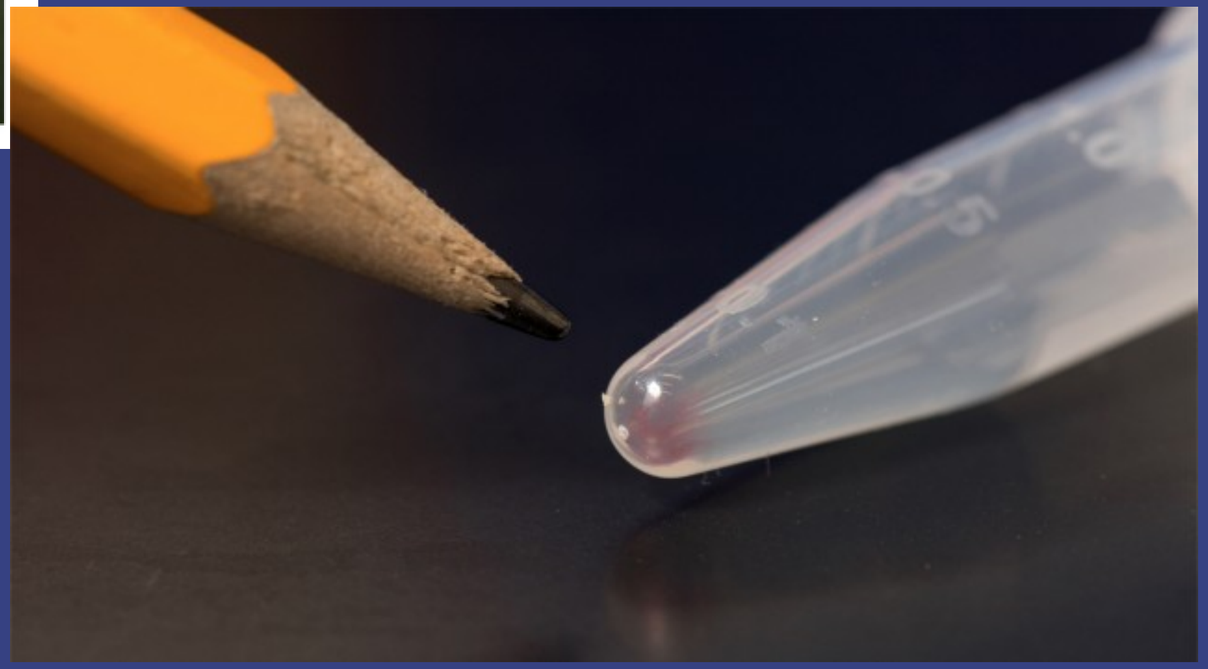

Memory Controller
Misc IO
Core
L1 cache
Core
Queue
Core
Core
Larger L2 cache
Shared L3 Cache
OPI 0
OPI 1

Microsoft and UW demonstrate first fully automated DNA data storage

https://www.youtube.com/watch?v=60Gi5lqL-dA



"Exabytes in a Test Tube: The Case for DNA Data Storage"



Warte śledzenia:

USENIX ATC '19

2019 USENIX
Annual Technical Conference

JULY 10–12, 2019
RENTON, WA, USA

www.usenix.org/atc19

DZIĘKUJĘ ZA UWAGĘ

EMAIL

jackowski@9livesdata.com

LINKEDIN

linkedin.com/in/ajackowski/

9LivesData