# معماری‌های نوظهور ذخیره‌ساز داده برای زیرساخت آتی فناوری اطلاعات

## Emerging Storage Architectures for future IT Infrastructure

Mostafa Kishani

Data Storage Systems, Networks, & Processing (DSN) Lab dsn.ce.sharif.edu
Dep. of Computer Engineering, Sharif University of Technology
HPDS Corporation www.hpds.ir

# کلیات ارائه

- معرفی شرکت پردازش و ذخیره‌سازی سریع داده (پرسا)
- معرفی سامانه‌های ذخیره‌سازی داده
- مشخصه‌های پیشرفته سامانه‌های ذخیره‌سازی داده
- معماری‌های نوظهور ذخیره‌ساز داده
- حافظه‌های نوظهور

# شرکت پردازش و ذخیره‌سازی سریع داده (پرسا)

**100**

\+

**200**

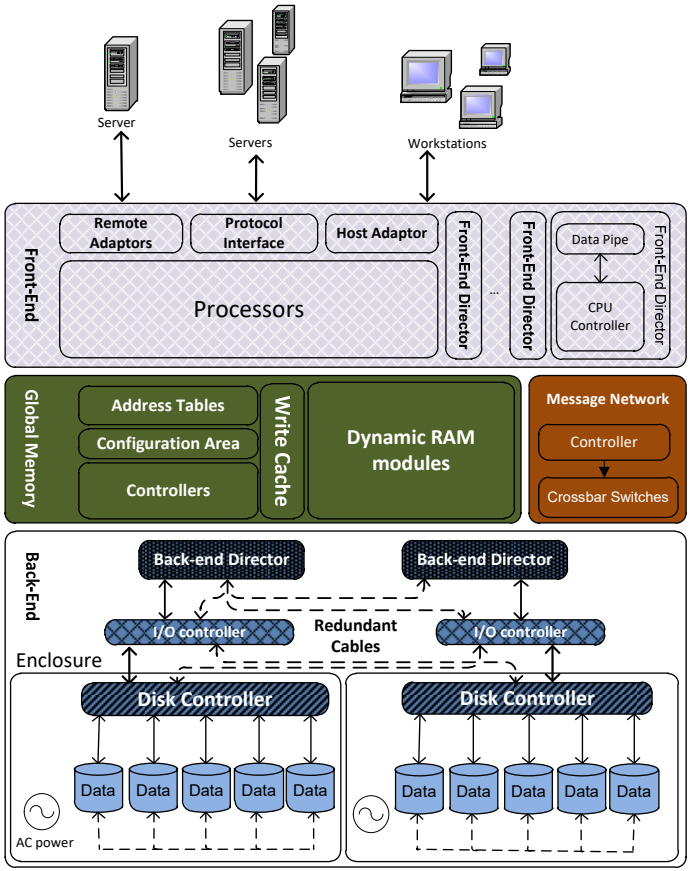
Indirect Labor


**Over 400 Products in the field**

**Over 10 Patents**


# سامانه‌های ذخیره‌سازی داده

- Data Storage Systems
  - High Capacity
  - High Available
  - High Performance




# محصولات HPDS

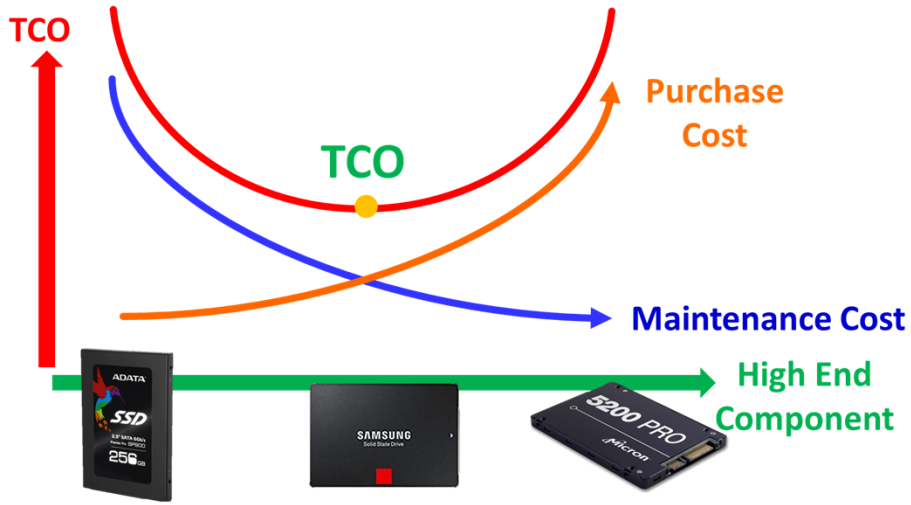
- SAB-SE (First Generation)
- SAB-HB (Hybrid)
- SAB ALL-FLASH
- SAB HB-V2 (Dual Board)

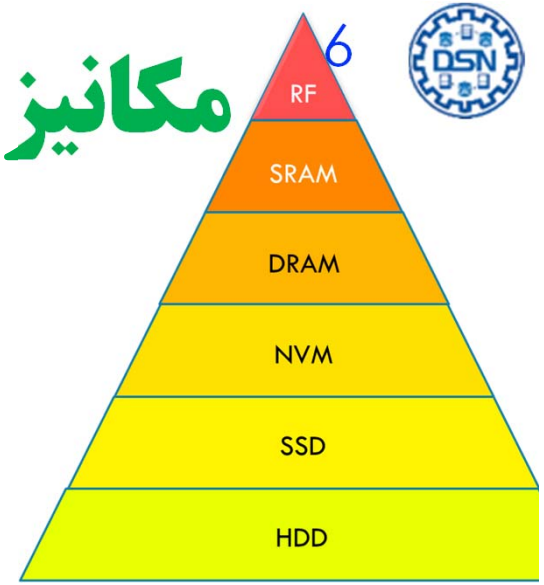



# مشخصه‌های پیشرفته سامانه‌های ذخیره‌سازی داده: چالش‌های علمی و فنی

# Caching مکانیزم‌های






## TWO-LEVEL CACHE

DRAM Cache

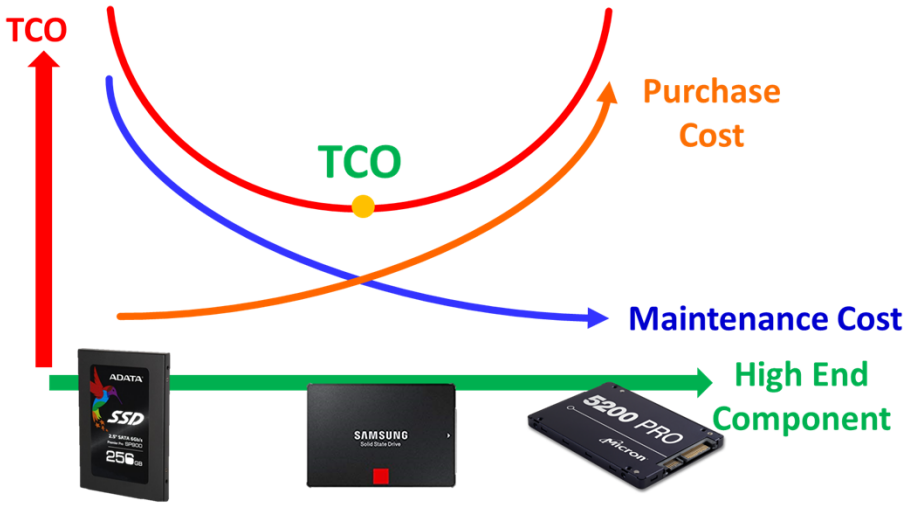
- Smart Caching Mechanism
- Optimized Cache Size
- Smart Device Purchase

Limited Endurance

SSD Cache


Disk

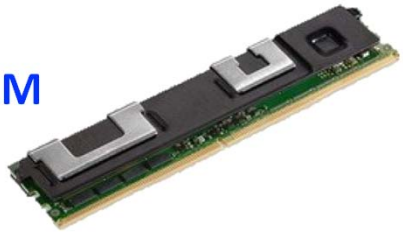

# مکانیزم‌های Caching




## THREE-LEVEL CACHE

DRAM Cache

NVM STTMRAM PCM ???



- Smart Caching Mechanism
- Optimized Cache Size
- Smart Device Purchase

Limited Endurance

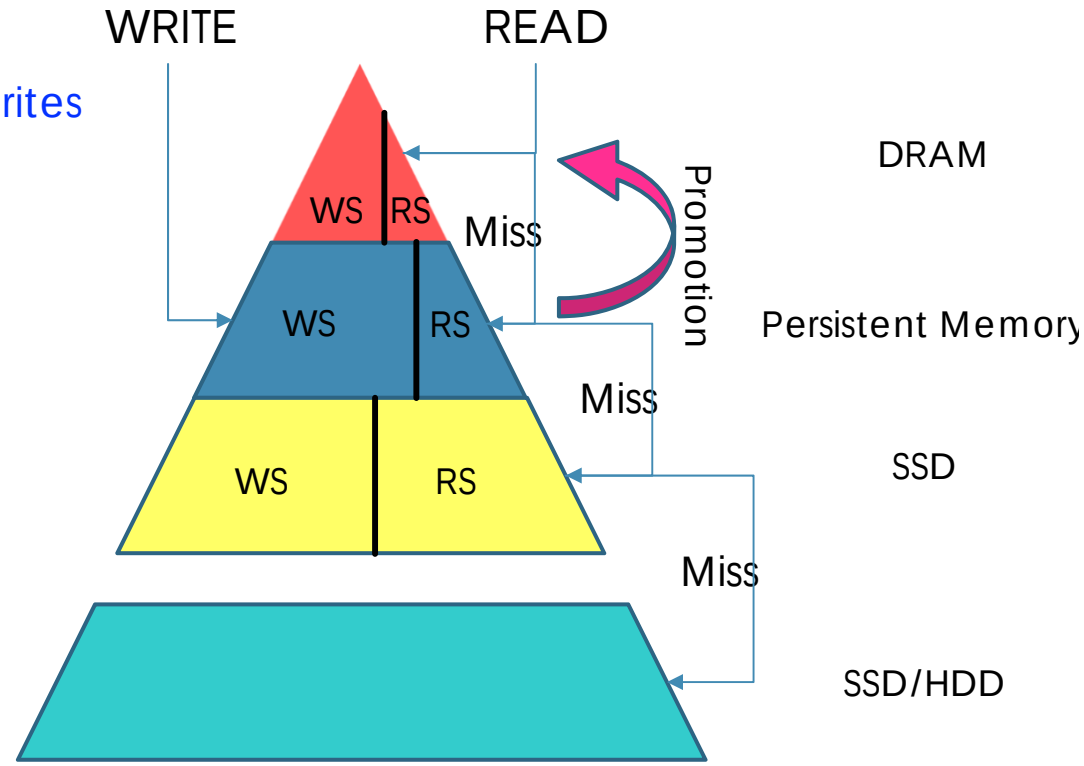
SSD Cache


Disk


# مکانیزم‌های Caching

For each cache layer:

- Write Share (WS):
  Share of cache dedicated to writes
- Read Share (RS): 1-WS

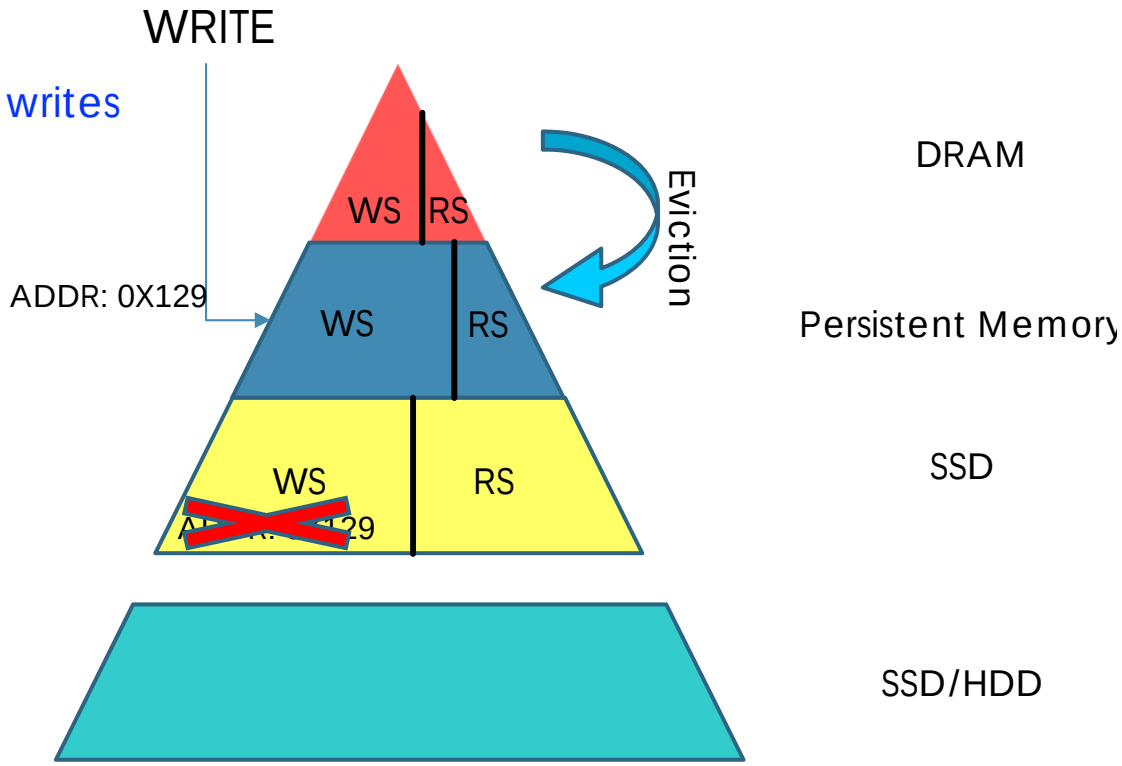


# مکانیزم‌های Caching



For each cache layer:

- Write Share (WS):
  Share of cache dedicated to writes
- Read Share (RS): 1-WS



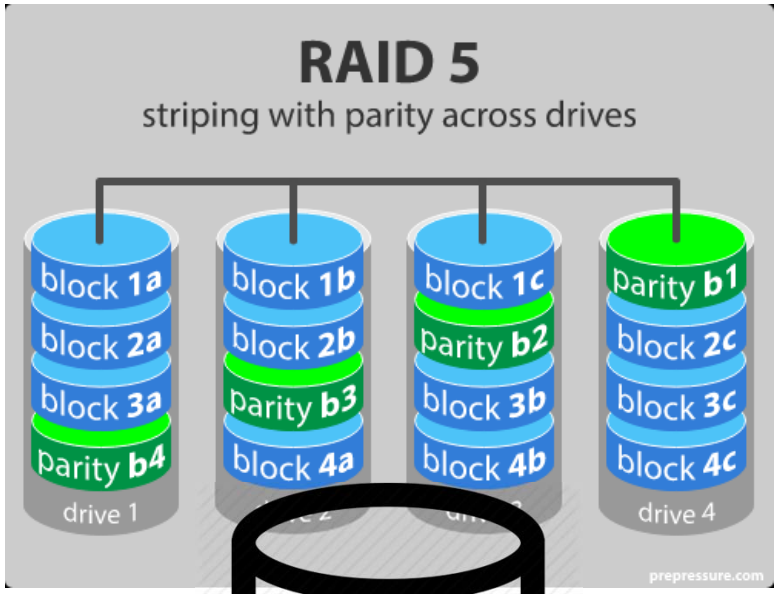
# معماری Dual-Board

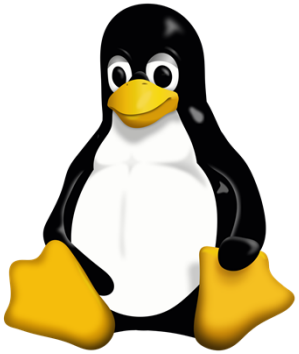
- Dual Board (Active/Active)
- Extremely High Available
  - 0.999999
- No Downtime at Update/Upgrade

- Automatic/Manual data migration between nodes
  - Manage Traffic
- Many Engineering Challenges:
  - Data Consistency
  - Node Synchronization
  - Dual Power Supply
  - Dual Chassis
  - Load Balancing
  - Failover
  - Split-Brain

# RAID نرم افزاری

- RAID: Redundant Array of Independent Disks

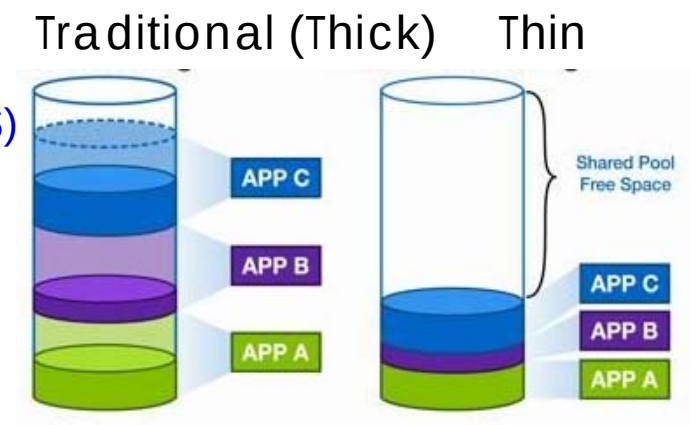


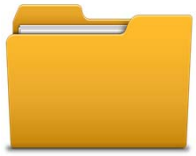
MD

# مشخصه‌های نرم‌افزاری پیشرفته

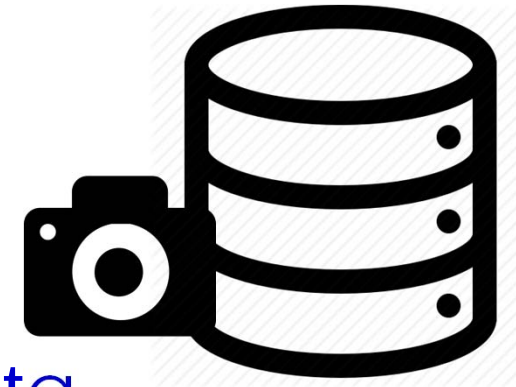
- Thin Provisioning




- Network Attached Storage (NAS)
  - Provides file service



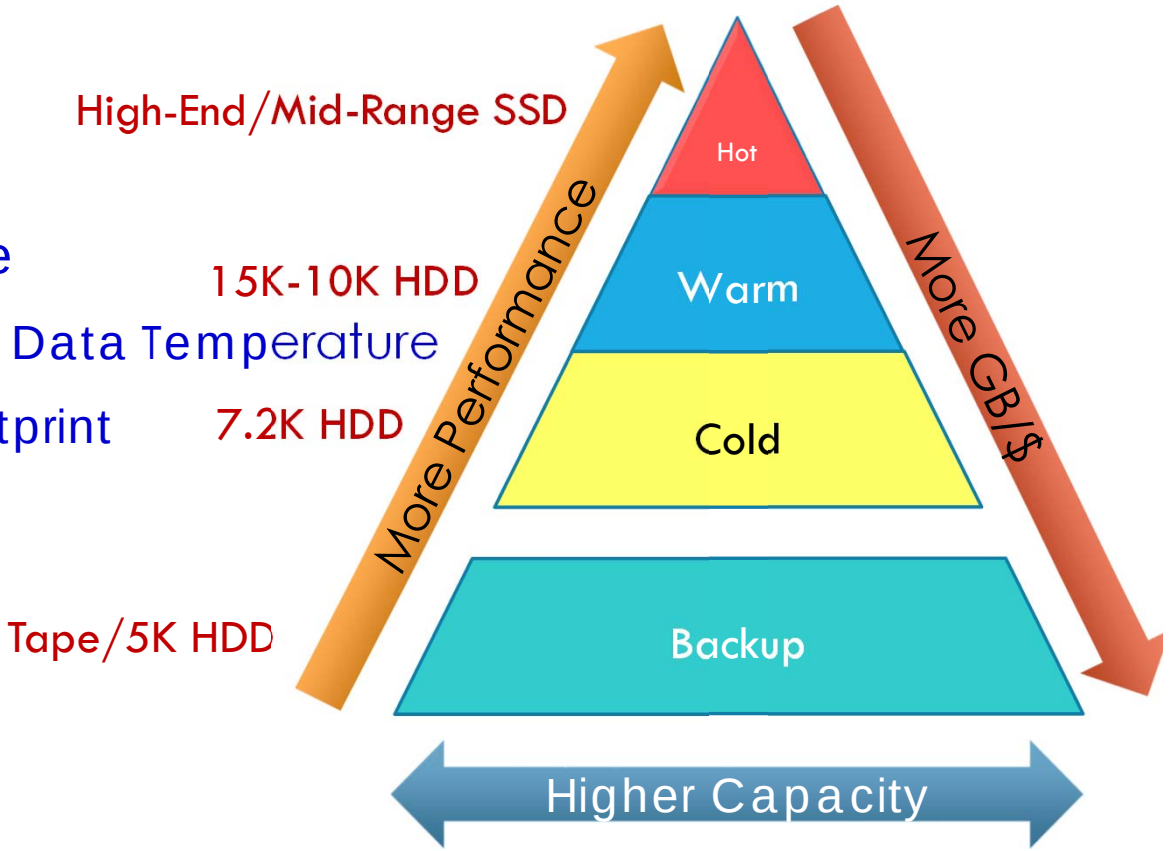
  - Support Linux/Win/ESXi

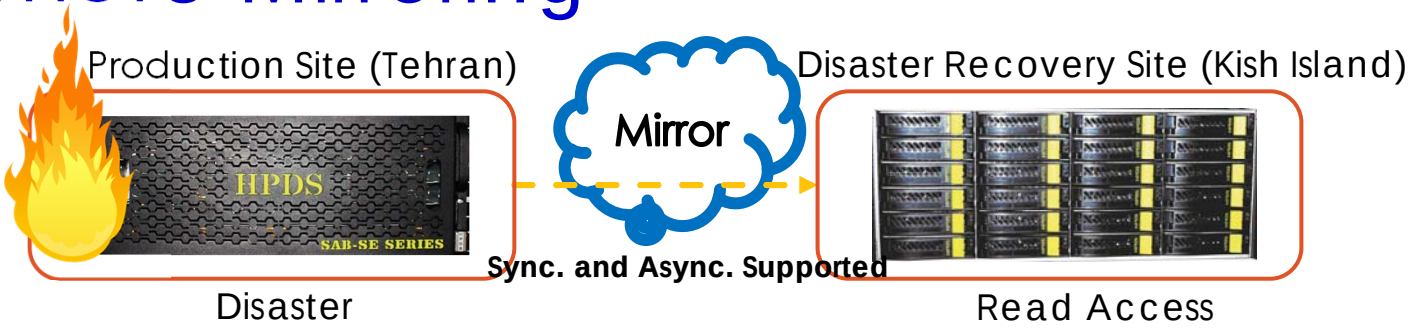
- Snapshot
  - Instantly takes an snapshot of data
  - Data can be recovered to snapshot anytime later

# مشخصه‌های نرم‌افزاری پیشرفته

- Tiering
  - Auto
    - Periodic Data Move
    - Automatically Detect Data Temperature
      - Using Access Footprint
  - Manual

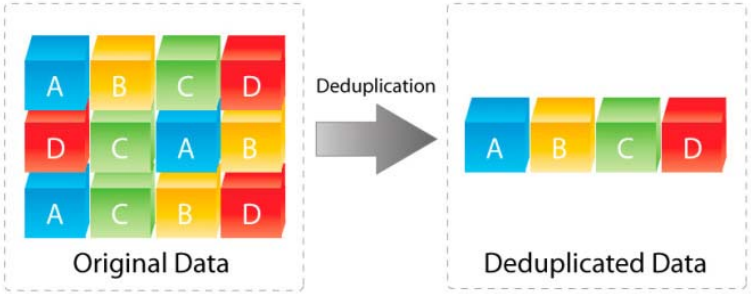


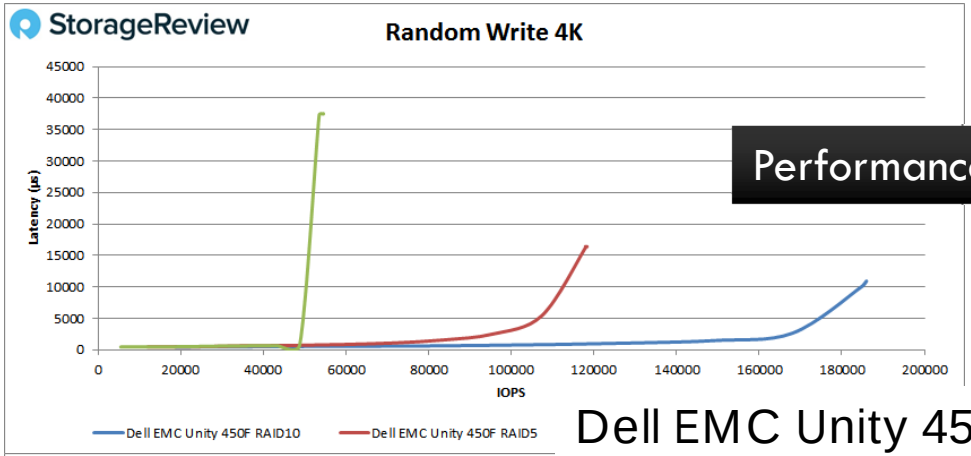

- Remote Mirroring



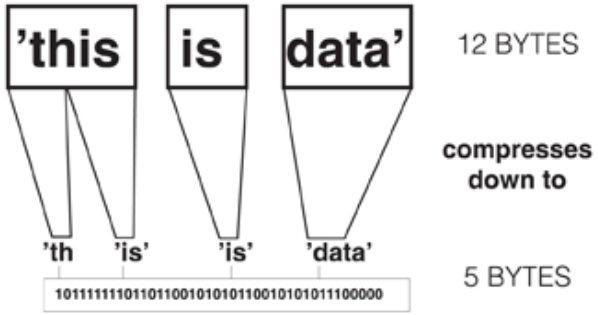
# مشخصه‌های نرم‌افزاری پیشرفته

- Deduplication
- Compression

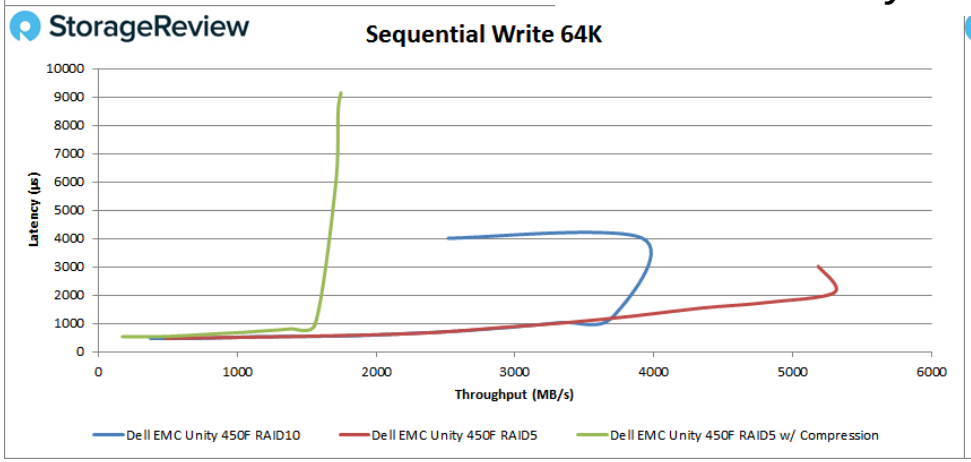


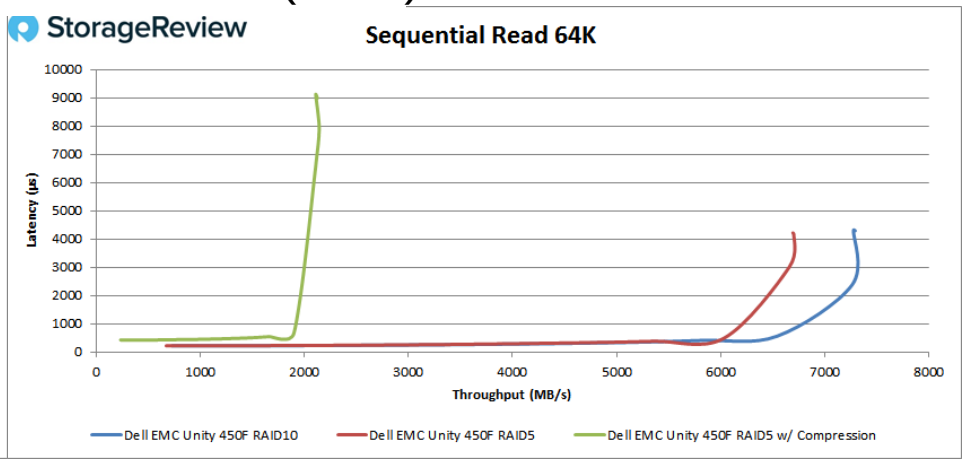




Performance Matters

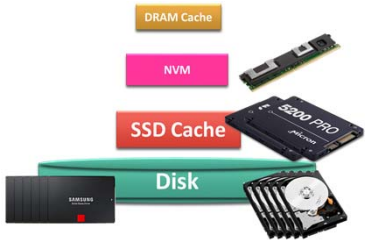



Dell EMC Unity 450F All-Flash (2018)

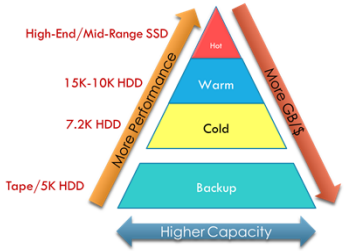


# موضوعات داغ تحقیقاتی




- مکانیزم‌های caching
  - چالش‌های اصلی: کارایی، هزینه، قابلیت اطمینان، Endurance




- الگوریتم‌های رده‌بندی داده
  - چالش‌های اصلی: کارایی، هزینه، Endurance
- الگوریتم‌های رمزنگاری، رمزگشایی، فشرده‌سازی و deduplication


  - چالش‌های اصلی: کارایی
- قابلیت اطمینان و دسترس‌پذیری
  - چالش‌های اصلی: حافظه‌های نوظهور، آرایه‌های دیسک، نرم‌افزار

# معماری‌های نوظهور ذخیره‌ساز داده

# معماری سنتی مراکز داده

# معماری همگرا

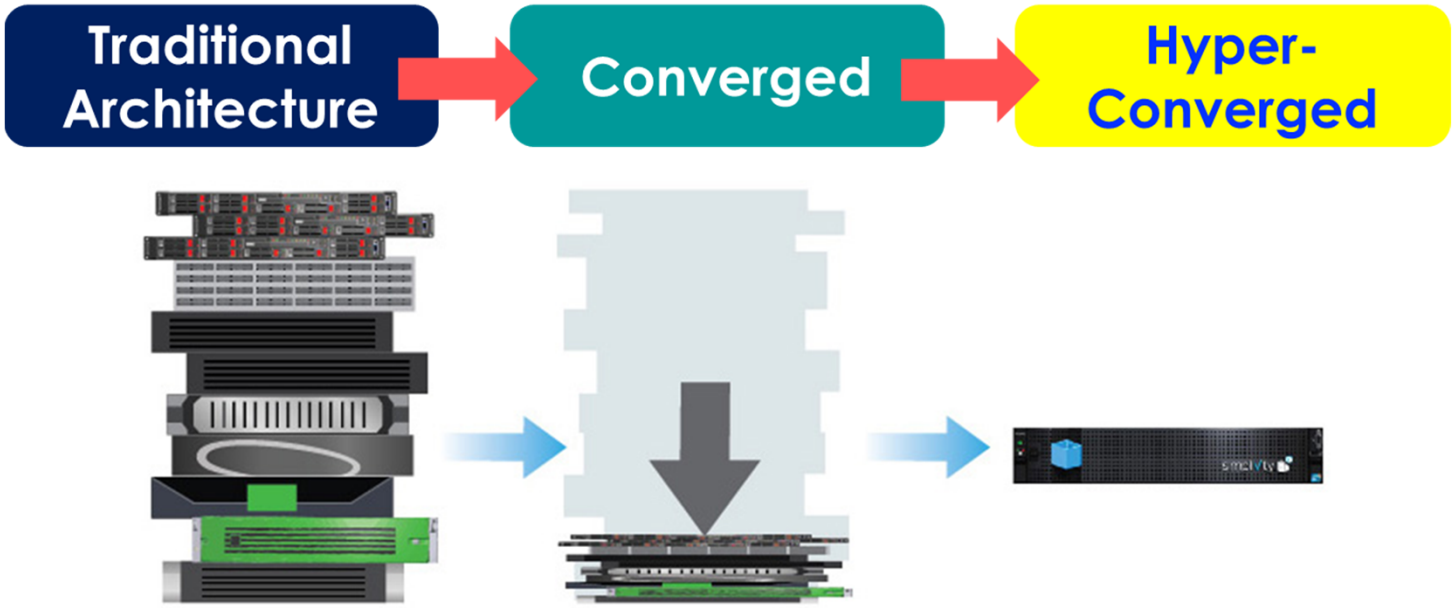
- Network, Storage, and Servers
  - Different Manufactures
  - Meet Industry Standards

1. Networking
2. Servers (Computing)
3. Storage

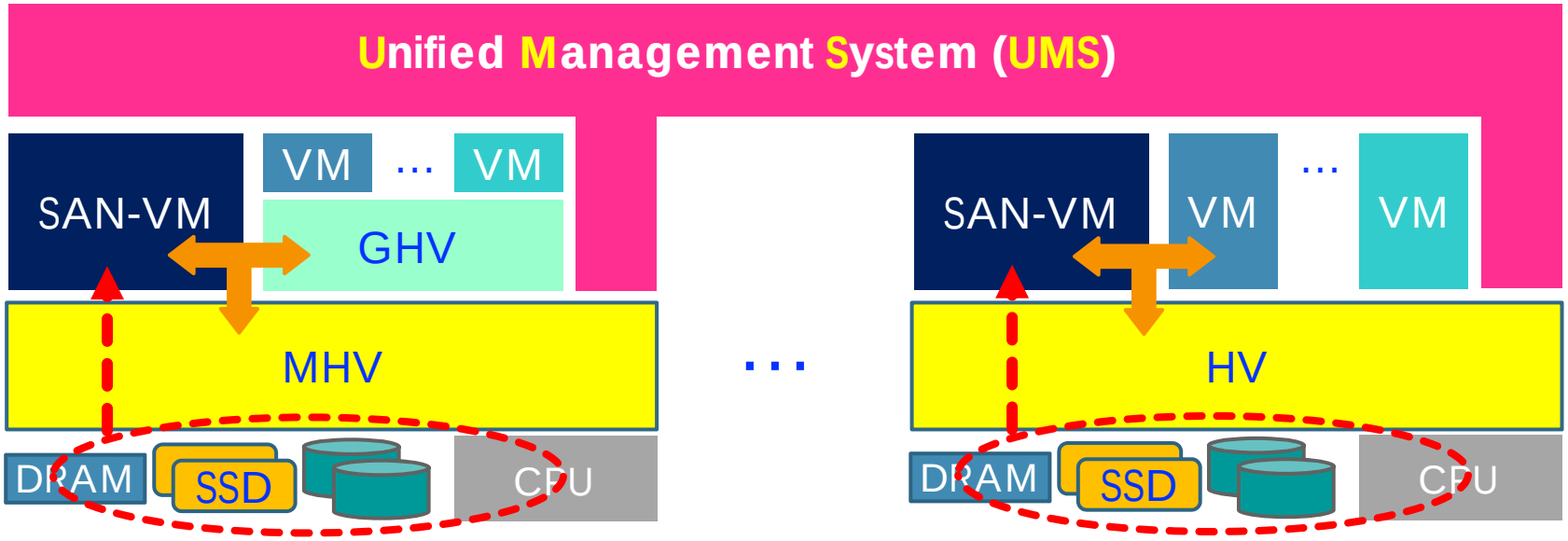
⚠ Expensive update
⚠ Risky data migration
⚠ Costly monitor and support
⚠ Long test and implementation timelines
⚠ Inefficient management
⚠ Lack of scalability

# زیرساخت‌های ابرهمگرا

- زیرساخت‌های ابرهمگرا
  - Hyper-Converged Infrastructures (HCI)
  - همگرایی و تجمیع زیرساخت‌های فناوری اطلاعات
  - مزایا
    - هزینه پایین، سادگی توسعه، مقیاس پذیری بالا
  - روند تکامل زیرساخت‌های فناوری اطلاعات



# معماری کلان

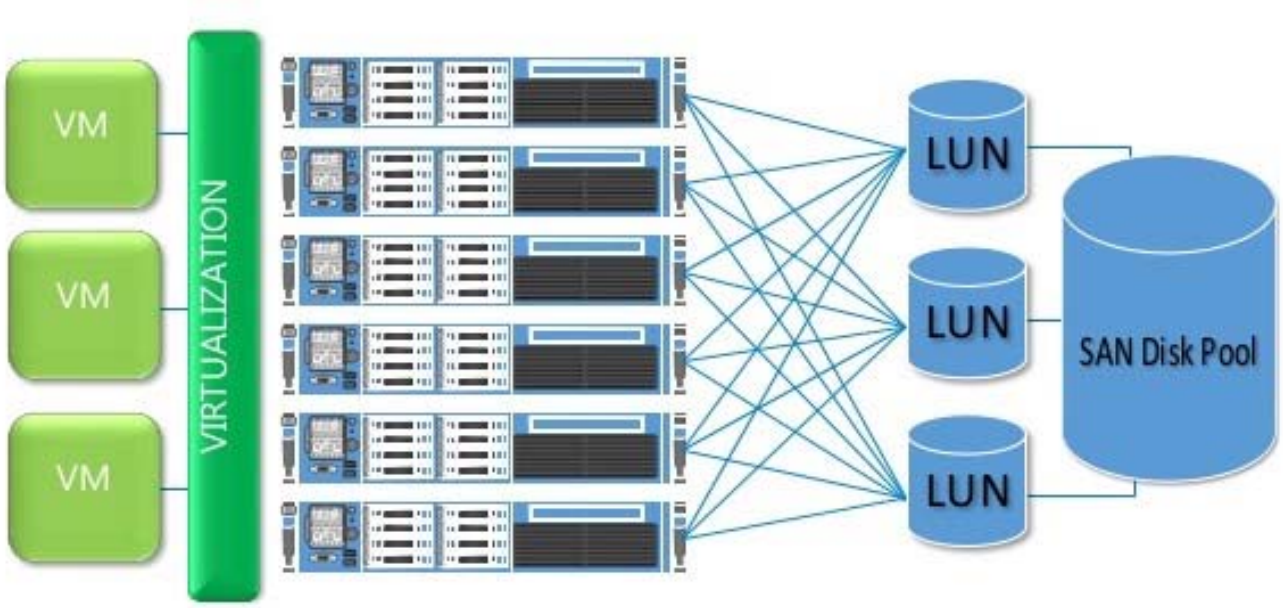



**Key Features:**

- ✓ COTS virtualization platform Hypervisor (HV)
- ✓ Unified Management System (UMS)
  - ✓ integrate the management of SAN and VMs in a unified environment
- ✓ Storage resources directly assigned to SAN-VM by HV
- ✓ Processing resources managed by HV
- ✓ SAN-VM and VMs connected via virtual network managed by HV

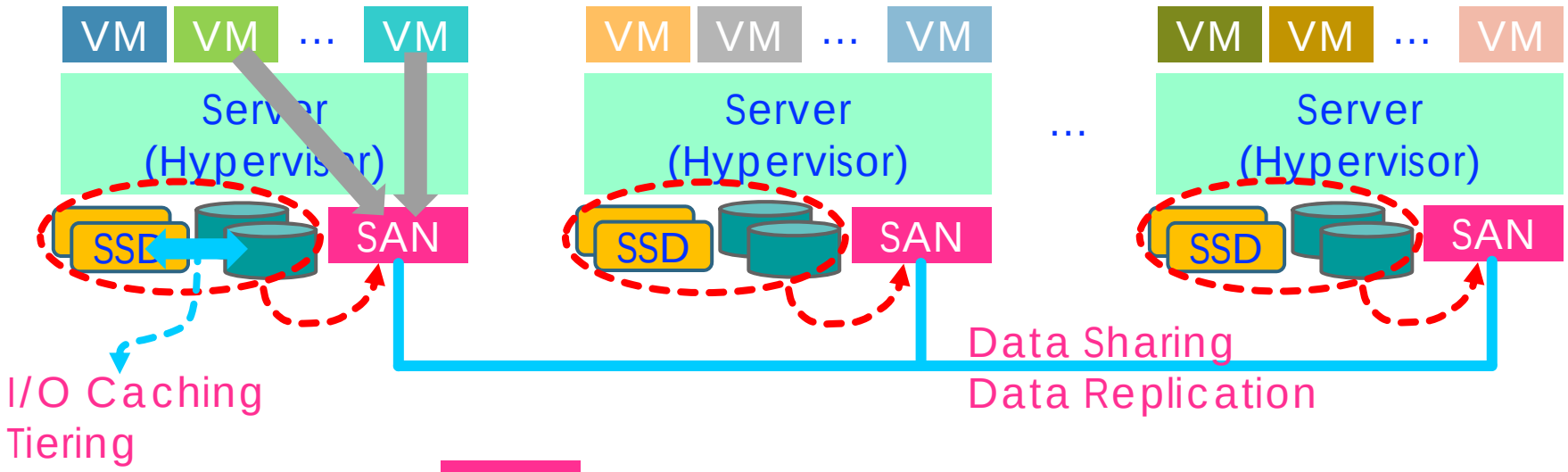
# معماری کلان (ادامه)

- Unified Management System



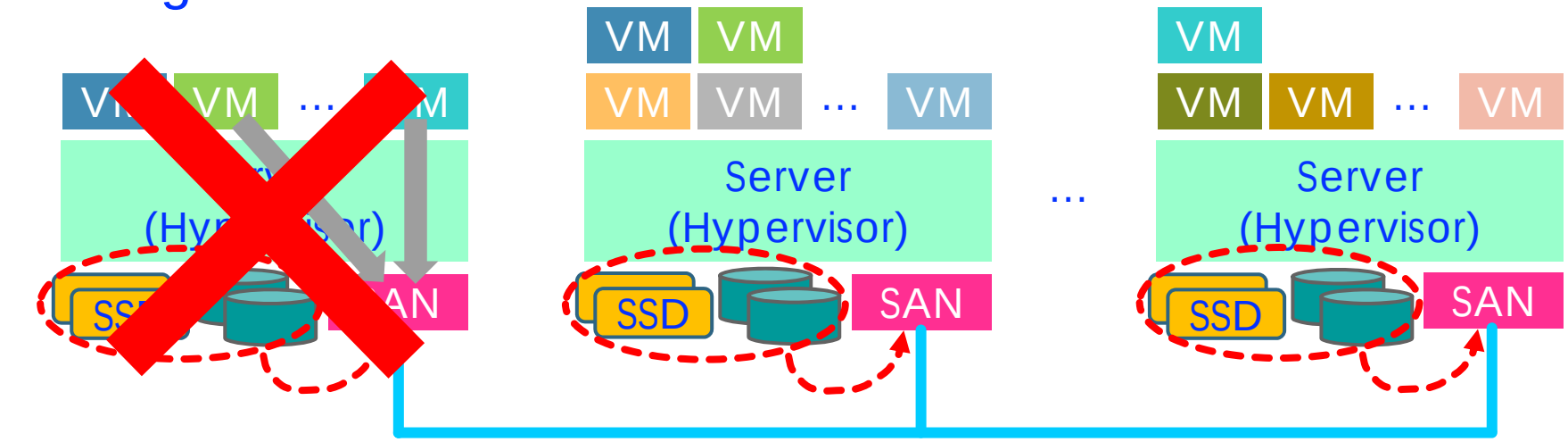

- Scale-Out Architecture
  - Reduced TCO
  - Low Provisioning Cost
  - High-Available

# دسترس‌پذیری و قابلیت اطمینان بالا

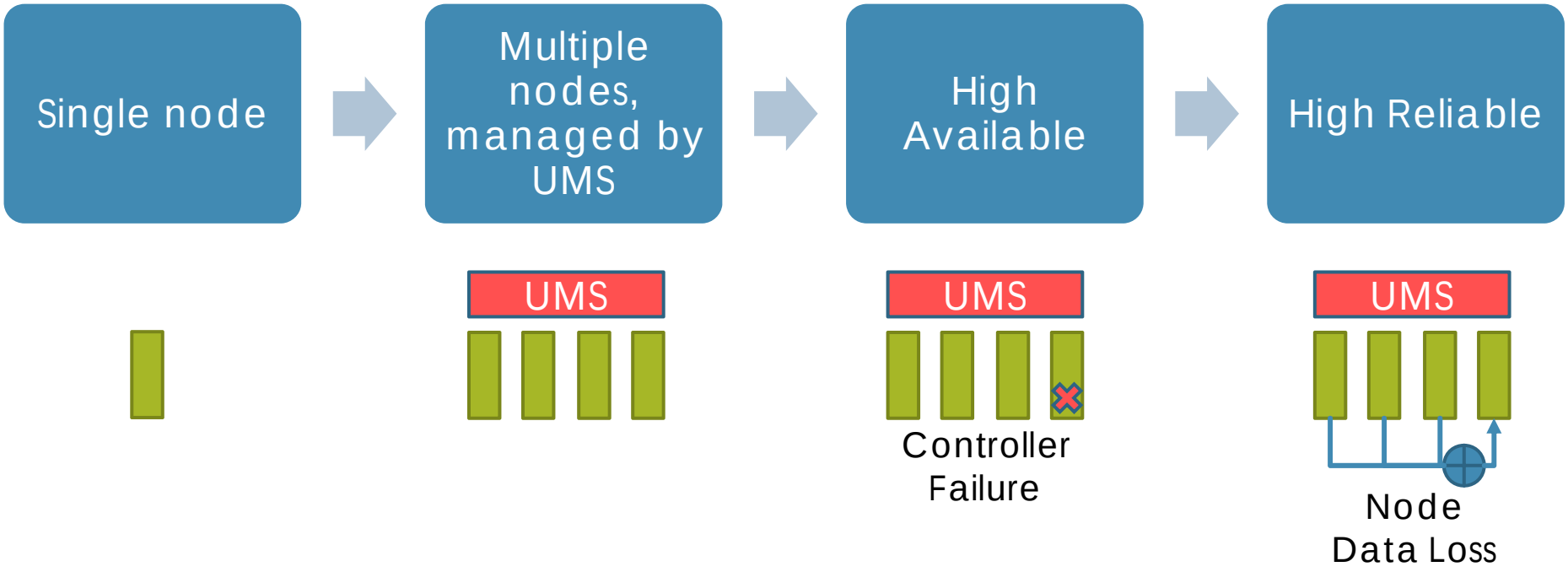



SAN Virtual Storage Controller
One virtual storage controller in each server

## Handling Node Failure



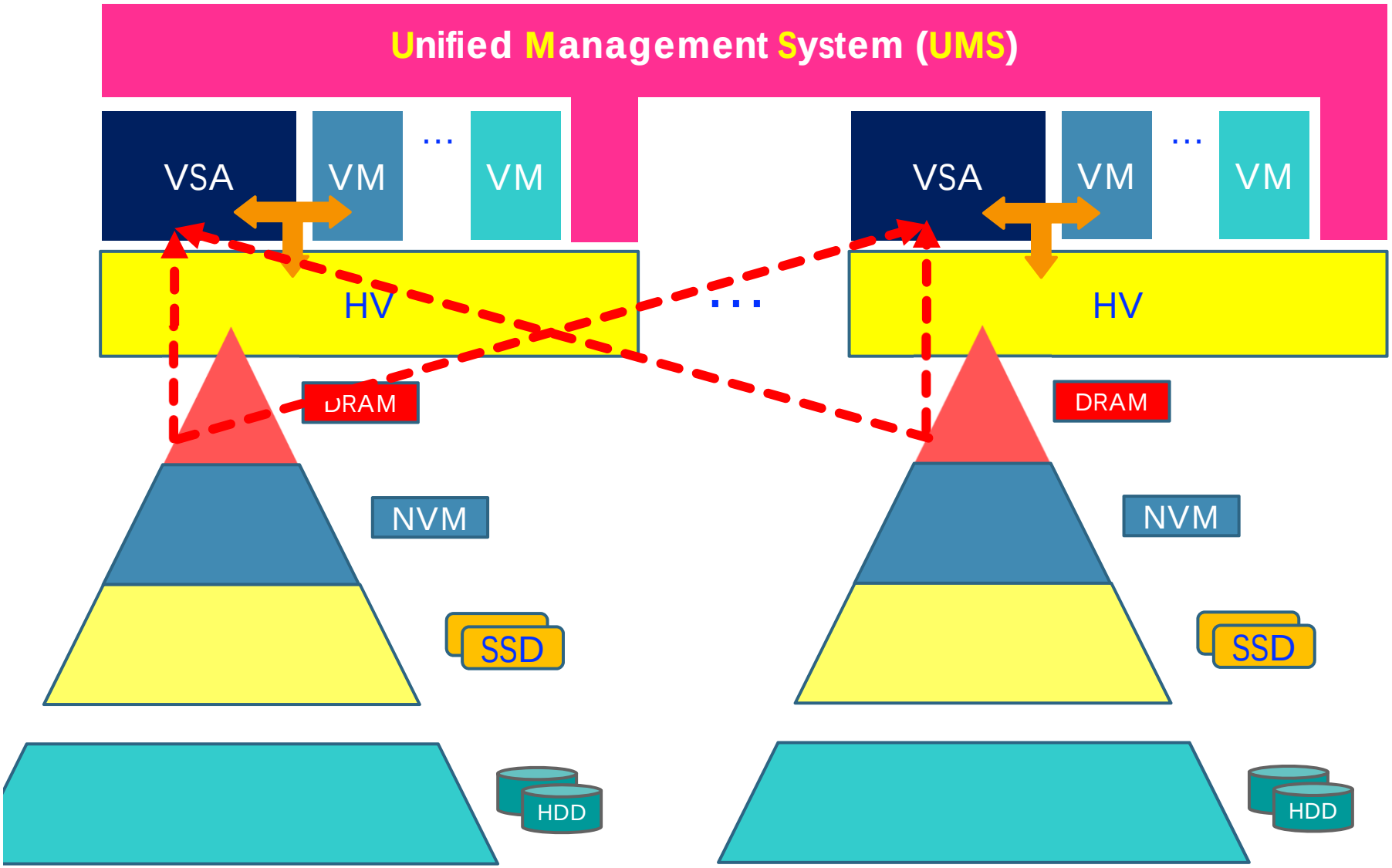
# مراحل بلوغ


Single node
Multiple nodes, managed by UMS
High Available
High Reliable
UMS
UMS
UMS
Controller Failure
Node Data Loss

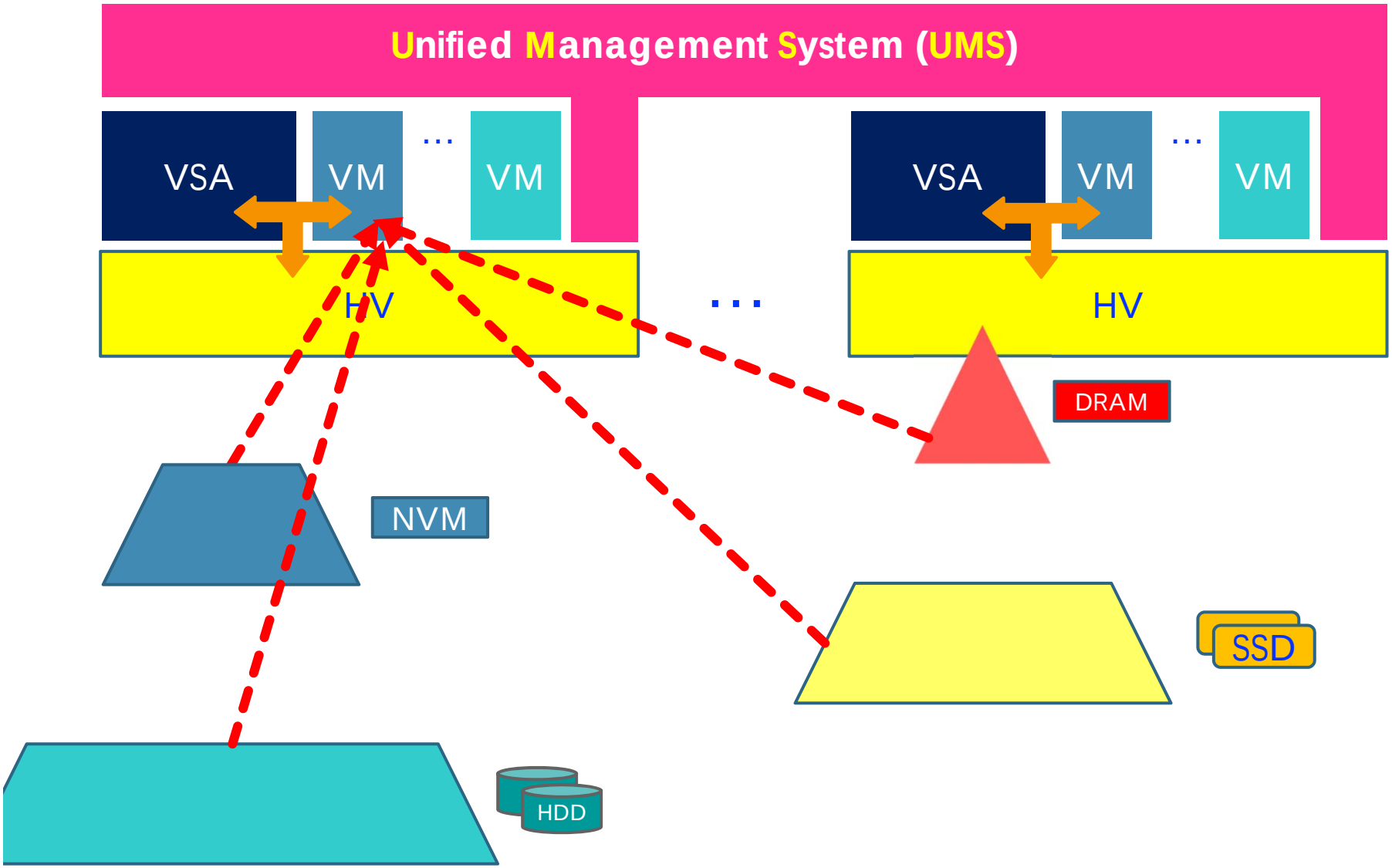
# موضوعات داغ تحقیقاتی

- قابلیت اطمینان و دسترس‌پذیری
  - توزیع داده بین گره‌ها
  - مکانیزم‌های HA
- مدیریت منابع
  - منابع پردازشی، DRAM، رده‌های رسانه ذخیره‌سازی، ترافیک شبکه
- مدیریت حافظه نهان
  - Load Balancing، کیفیت سرویس
- کارایی
  - تاخیر دسترسی، Throughput

# مدیریت منابع و حافظه نهان

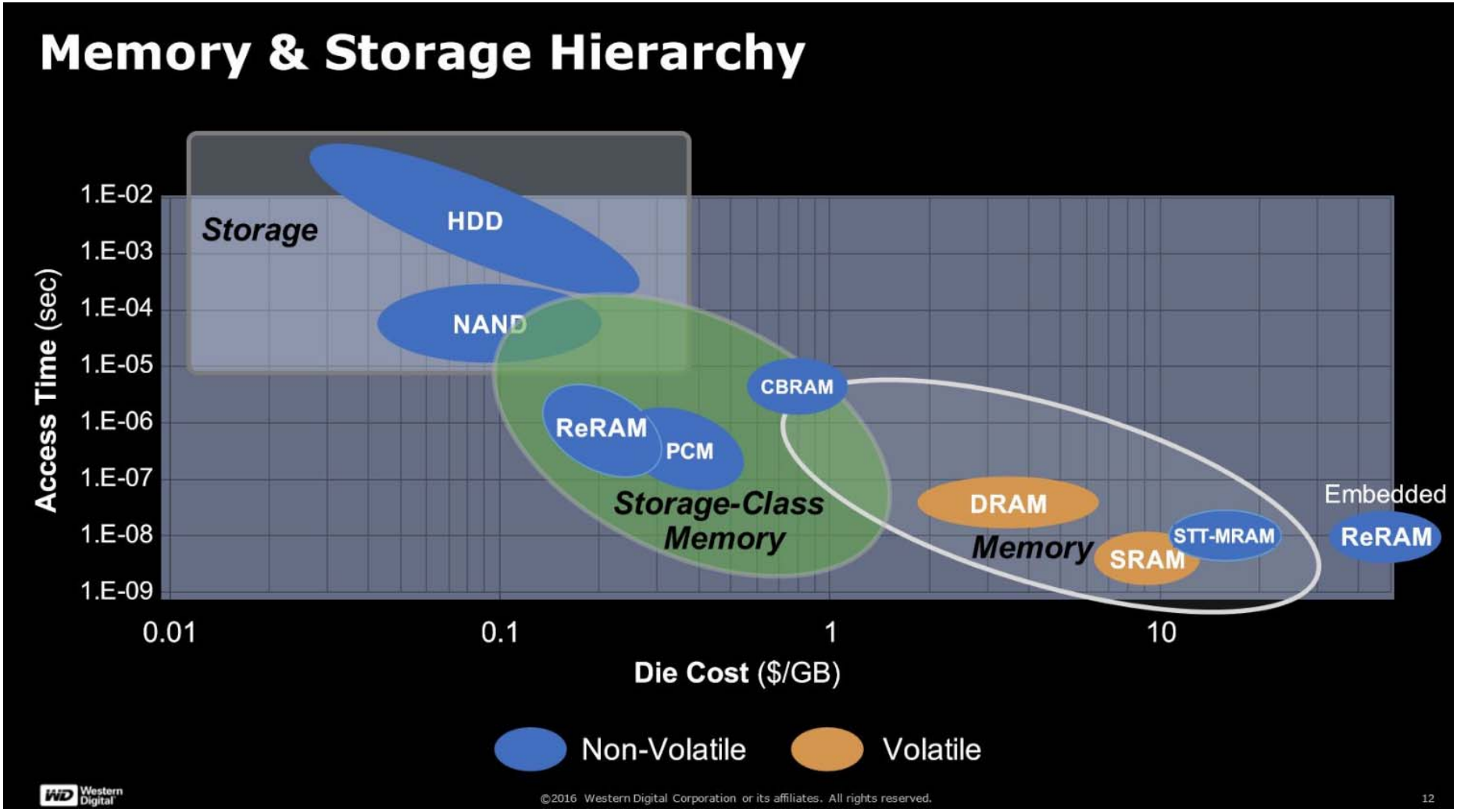


# مدیریت منابع و حافظه نهان



# حافظه‌های نوظهور

# حافظه‌های نوظهور



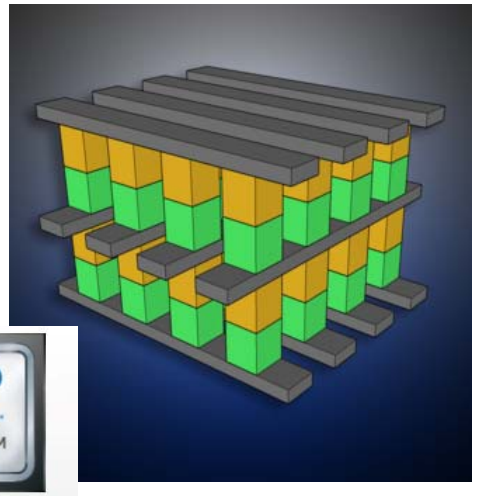
# حافظه‌های نوظهور

| | Memristor | PCM | STT-RAM | ReRAM | DRAM | Flash | HDD |
|---|---|---|---|---|---|---|---|
| **Read Time (ns)** | <10 | 20-70 | 10-30 | 10 | 10-50 | 25,000 | $5\text{-}8\text{x}10^6$ |
| **Write Time (ns)** | 20-30 | 50-500 | 13-95 | 1-100 | 10-50 | 200,000 | $5\text{-}8\text{x}10^6$ |
| **Endurance (cycles)** | 1 Trillion | 10 - 100 million | $10^{15}$ | $10^{10}$-$10^{12}$ | >$10^{17}$ | 500-$10^6$ | $10^{15}$ |
| **Retention (without power)** | >10 Years | <10 Years | Weeks | Months | <Second | ~10 Years | ~10 Years |
| **Energy Per Bit (pj)[2]** | 0.1-3 | 2-100 | 0.1-1 | ? | 2-4 | $10^1$-$10^{14}$ | $10^6$-$10^7$ |
| **Chip Area Per Bit ($F^2$)** | 4 | 8-16 | 14-64 | ? | 6-8 | 4-8 | n/a |

# بازار SCM

- 3D Xpoint
  - Inherits PCM and ReRAM

IMFLASH an intel, micron venture



intel OPTANE DC PERSISTENT MEMORY



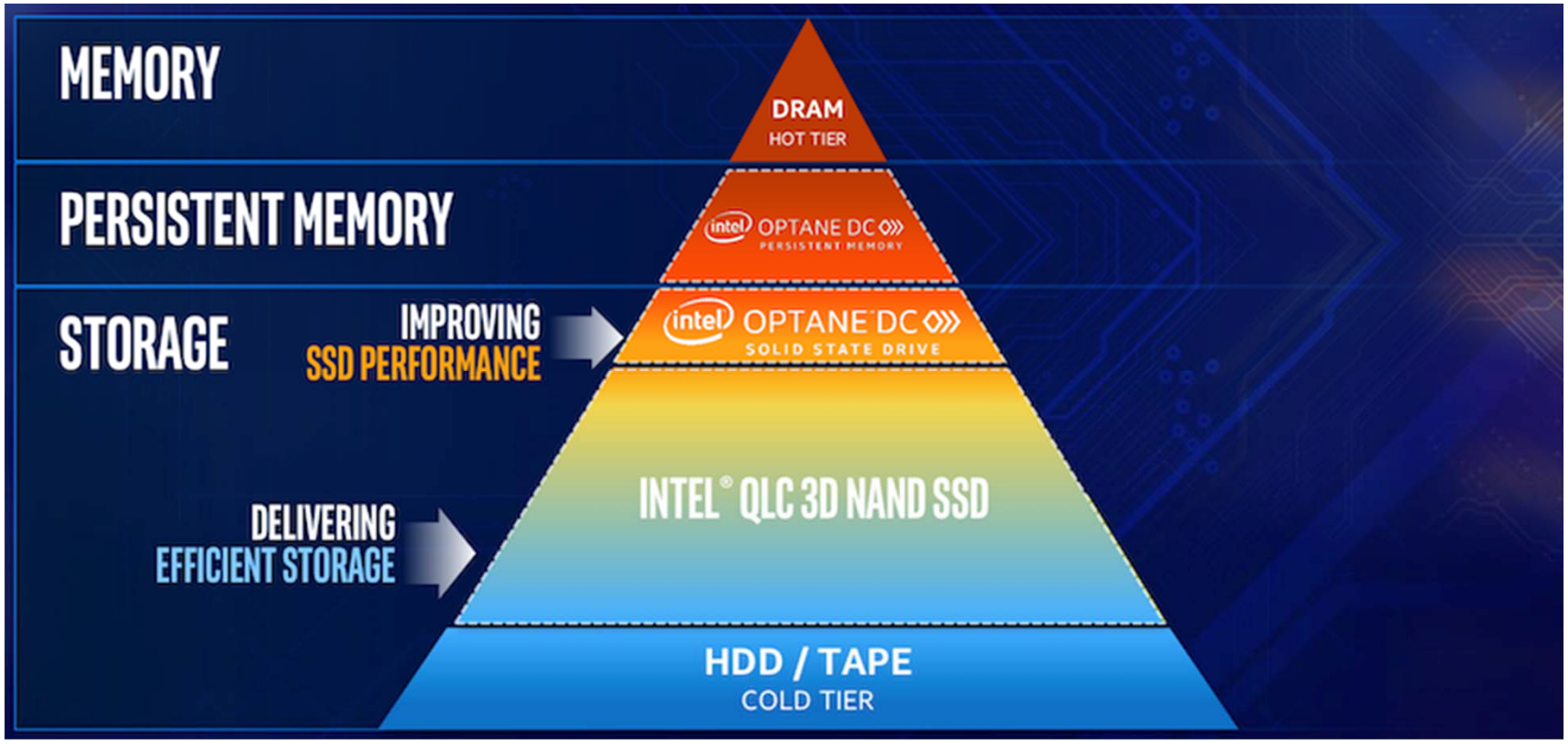
- 3D ReRAM
  - Not expected before 2020

WD Western Digital®

TOSHIBA

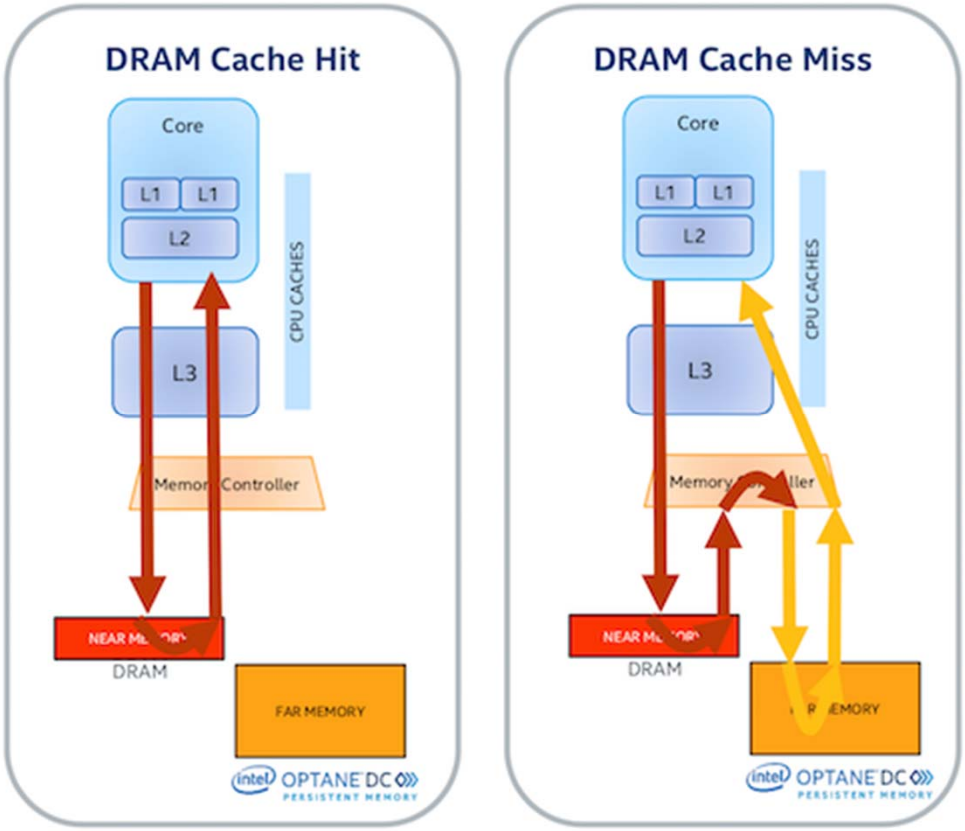
# Intel Optane DC Persistent Memory Module (PMM)

- Fills the gap between DRAM and NVM



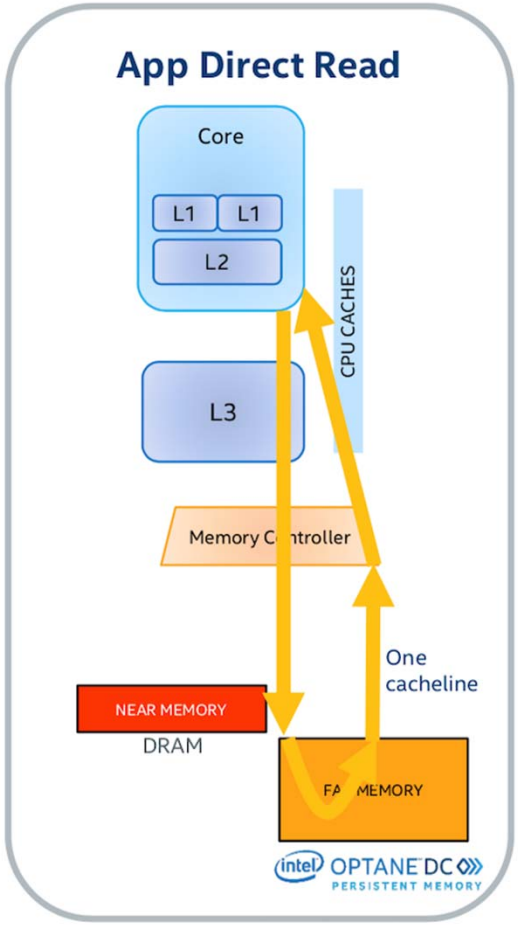
# Intel Optane DC PMM

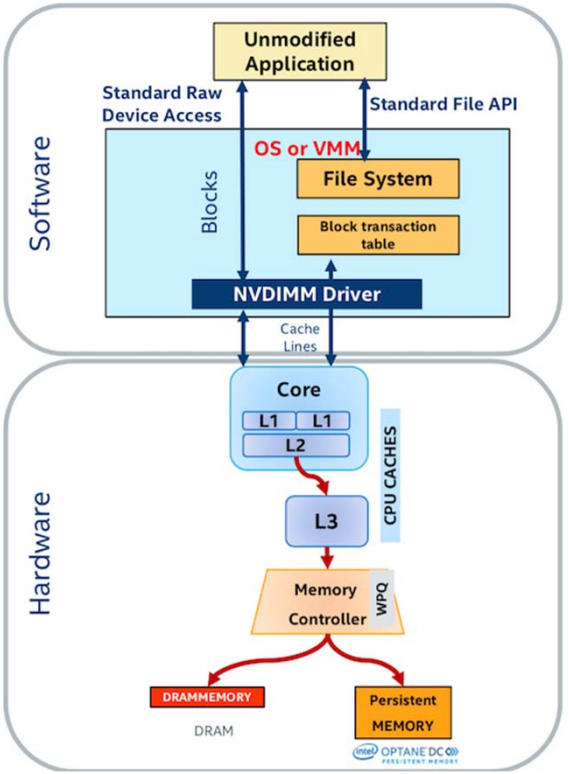
- Memory Mode



# Intel Optane DC PMM

- App Direct Mode



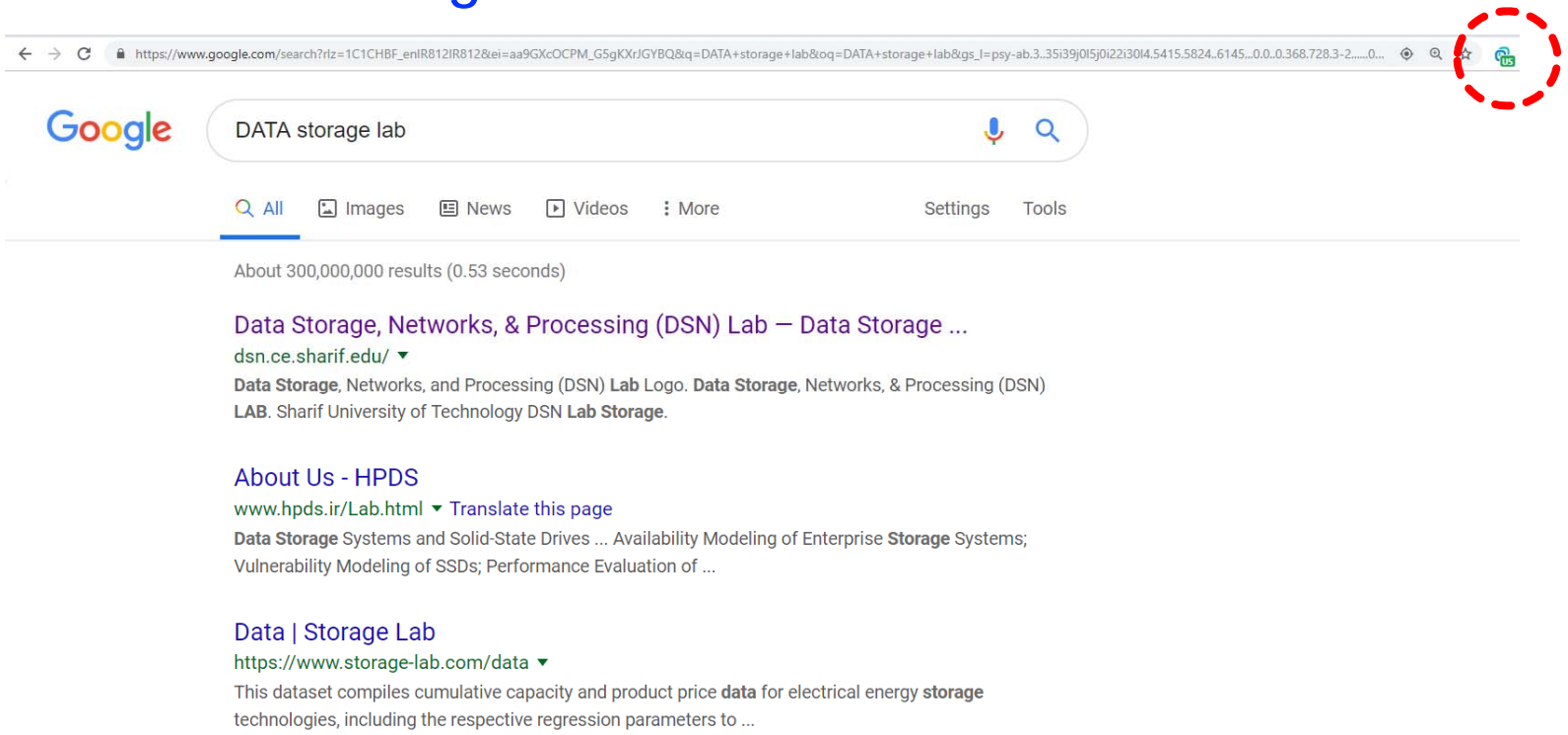



# آزمایشگاه تحقیقاتی ذخیره‌سازی، پردازش و شبکه‌های داده (DSN)

# آزمایشگاه DSN

- dsn.ce.sharif.edu
  - Just google it
    - Data storage laboratory
    - Data storage lab

Data Storage, Networks, & Processing (DSN) LAB



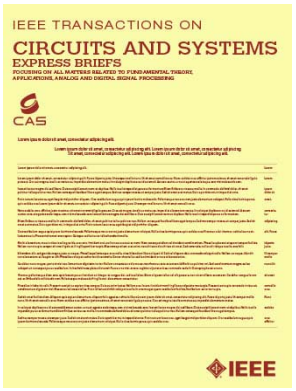

- **Patents**
  - H. Asadi and S. Ahmadian, “Cache Allocation to a Virtual Machine,” **US Patent**, Application Pending, App. No. 16252584, Jan. **2019**.
  - H. Asadi, R. Salkhordeh, and S. Ebrahimi, “Reconfigurable Caching”, **US Patent**, Application Pending, App. No. 16243298, Jan. **2019**.
  - H. Asadi, Z. Ebrahimi, and B. Khaleghi, “Programmable Logic Design,” **US Patent**, No. 10,312,918, Filed: Feb. 13, 2017, Granted: June 04, **2019**.
  - H. Asadi, R. Salkhordeh, and S. Ebrahimi, “OS-Level Data Tiering to Improve Performance of RAID Arrays”, Iran State Organization for Deeds and Properties, Application No. 13945014000300293 7, Approved, Feb. 2017.
  - H. Asadi, R. Salkhordeh, and S. Ebrahimi, “Re-configurable I/O Caching Architecture with Online Workload Characterization”, Iran State Organization for Deeds and Properties, Pending, 2016.

# Recent Publications

## 2018-2019: Journal Papers

- IEEE Transactions on Reliability (TR): 3
- IEEE Transactions on Computers (TC): 3
- IEEE Transactions on Parallel & Distributed Systems (TPDS): 2
- IEEE Transactions on VLSI (TVLSI): 1
- IEEE Transactions on CAD (TCAD): 1
- IEEE Transactions on Circuits and Systems I (TCAS-I): 1








- Recent Publications
  - 2017-2019: Conference Papers
    - IEEE/ACM Design, Automation and Test in Europe Conference (DATE): 5
      - DATA 2019 BEST PAPER AWARD
    - ACM SIGMETRICS: 1
    - ASP-DAC: 1

# آزمایشگاه DSN


- E. Cheshmikhani, H. Farbeh, and H. Asadi, “A System-Level Framework for Analytical and Empirical Reliability Exploration of STT-MRAM Caches", IEEE Transactions on Reliability (**TR**), In Press, 2019.
- R. Salkhordeh, O. Mutlu, and H. Asadi, “An Analytical Model for Performance and Lifetime Estimation of Hybrid DRAM-NVM Main Memories,” IEEE Transactions on Computers (**TC**), Vol. 68, Issue 8, August 2019.
- R. Salkhordeh, M. Hadizadeh, and H. Asadi, “An Efficient Hybrid I/O Caching Architecture Using Heterogeneous SSDs,” IEEE Transactions on Parallel & Distributed Systems (**TPDS**), Vol. 30, Issue 6, June 2019.
- M. Kishani, M. B. Tahoori, and H. Asadi, “Dependability Analysis of Data Storage Systems in Presence of Soft Errors”, IEEE Transactions on Reliability (**TR**), Vol. 68, Issue 1, March 2019.
- B. Khaleghi, B. Omidi, H. Amrouch, J. Henkel, and H. Asadi, “Estimating and Mitigating Aging Effects in Routing Network of FPGAs,” IEEE Transactions on VLSI (**TVLSI**), Vol. 27, Issue 3, March 2019.
- E. Cheshmikhani, H. Farbeh, S.G. Miremadi, and H. Asadi, “TA-LRW: A Replacement Policy for Error Rate Reduction in STT-MRAM Caches”, IEEE Transactions on Computers (**TC**), Vol. 68, No. 3, March 2019.
- S. Tamimi, Z. Ebrahimi, B. Khaleghi, and H. Asadi, “An Efficient SRAM-based Reconfigurable Architecture for Embedded Processors”, IEEE Transactions on CAD (**TCAD**), Vol. 38, Issue 3, March 2019.
- M. Kishani and H. Asadi, “Modeling Impact of Human Errors on the Data Unavailability and Data Loss of Storage Systems”, IEEE Transactions on Reliability (**TR**), Vol. 67, Issue 3, Sept. 2018.
- B. Khaleghi and H. Asadi, "A Resistive RAM-Based FPGA Architecture Equipped with Efficient Programming Circuitry”, IEEE Transactions on Circuits and Systems I: Regular Papers (**TCAS-I**), Vol. 65, Issue 7, July 2018.
- S. Ahmadian, O. Mutlu, and H. Asadi, “ECI-Cache: A High-Endurance and Cost-Efficient I/O Caching Scheme for Virtualized Platforms”, Proceedings of the ACM on Measurement and Analysis of Computing Systems, Vol. 2, No. 1, March 2018.
- R. Salkhordeh, S. Ebrahimi, and H. Asadi, “ReCA: an Efficient Reconfigurable Cache Architecture for Storage Systems with Online Workload Characterization”, IEEE Transactions on Parallel & Distributed Systems (**TPDS**), Vol. 29, Issue 7, July 2018.

# آزمایشگاه DSN


- S. Ahmadian, R. Salkhordeh, and H. Asadi, “LBICA: A Load Balancer for I/O Cache Architectures”, IEEE/ACM Design, Automation and Test in Europe Conference (**DATE**), Florence, Italy, March 2019.
- E. Cheshmikhani, H. Farbeh, and H. Asadi, “Enhancing Reliability of STT-MRAM Caches by Eliminating Read Disturbance Accumulation”, IEEE/ACM Design, Automation and Test in Europe Conference (**DATE**), Florence, Italy, March 2019 (Best Paper Award).
- E. Cheshmikhani, H. Farbeh, and H. Asadi, “ROBIN: Incremental Oblique Interleaved ECC for Reliability Improvement in STT-MRAM Caches”, 24th IEEE Asia and South Pacific Design Automation Conference (**ASP-DAC**), Tokyo, Japan, Jan. 2019.
- S. Ahmadian, O. Mutlu, and H. Asadi, “ECI-Cache: A High-Endurance and Cost-Efficient I/O Caching Scheme for Virtualized Platforms”, **ACM SIGMETRICS**, Irvine, California, June 2018 (*Proceedings will be published in a special issue by ACM POMACS; **Also listed in the journal papers).
- S. Ahmadian, F. Taheri, M. Lotfi, M. Karimi, and H. Asadi, “Investigating Power Outage Effects on Reliability of Solid-State Drives”, IEEE/ACM Design, Automation and Test in Europe Conference (**DATE**), Dresden, Germany, March 2018.
- Z. Seifoori, B. Khaleghi, and H. Asadi, "A Power Gating Switch Box Architecture in Routing Network of SRAM-Based FPGAs in Dark Silicon Era", IEEE/ACM Design, Automation and Test in Europe Conference (**DATE**), Lausanne, Switzerland, March 2017.
- M. Kishani, R. Eftekhari, and H. Asadi, "Evaluating Impact of Human Errors on the Availability of Data Storage Systems", IEEE/ACM Design, Automation and Test in Europe Conference (**DATE**), Lausanne, Switzerland, March 2017.

# از توجه شما سپاسگزارم