Disaggregated storage and role of storage stack

Manoj Wadekar

Principal Architect

eBay Marketplace at a Glance

One of the world's largest and most vibrant marketplaces

$22.5B
GMV

1.1B
Live listings

80%
Items sold as new

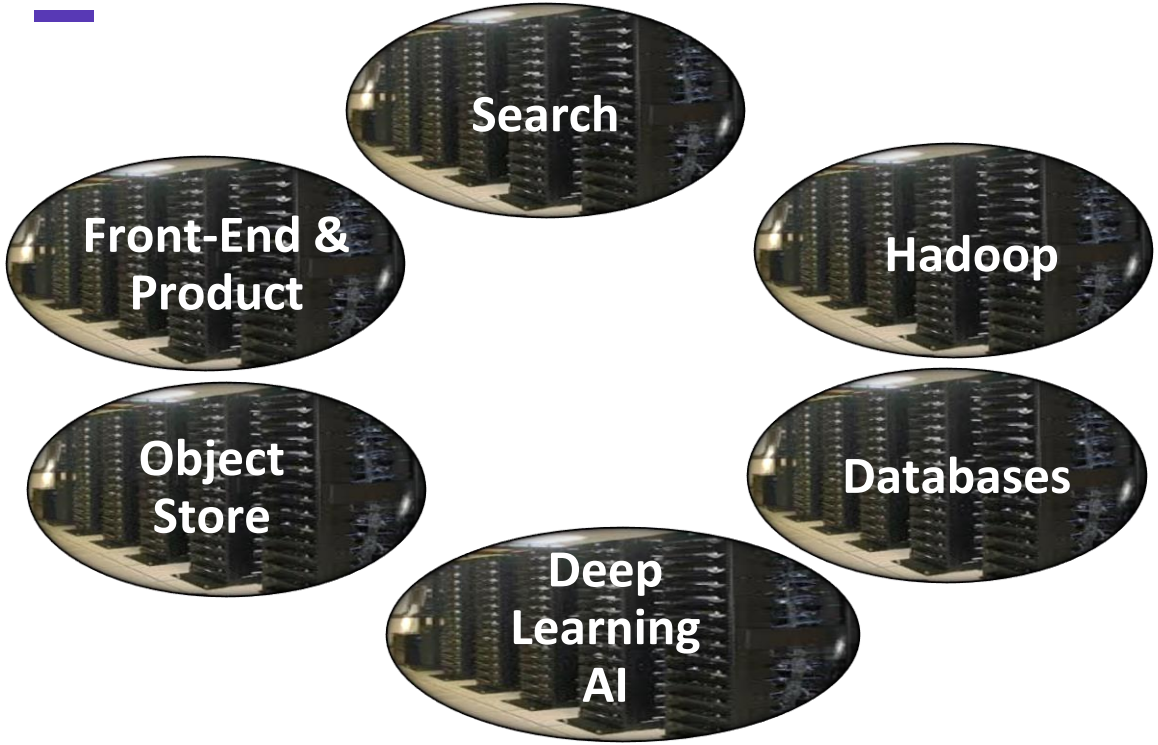
70%
Transactions that ship for free (US, UK, DE)

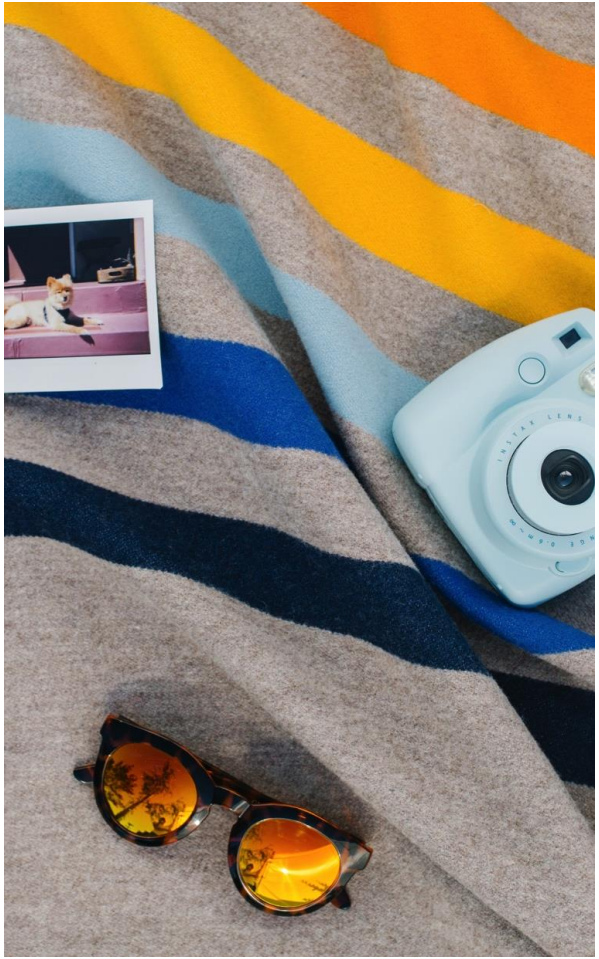
13.7M
New listings added via mobile per week

62%
Platform GMV touched by mobile

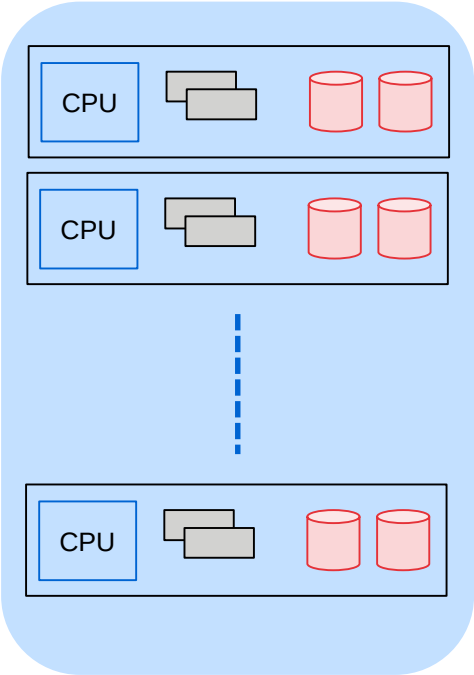
Data as of 12/31/2017

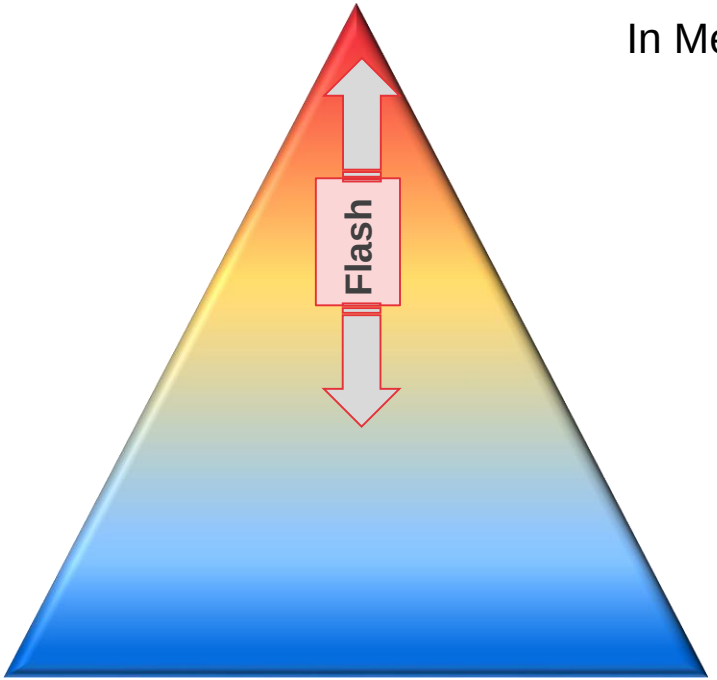
ebay Hyper scale Infrastructure





Typical hyper scale servers





Performance Needs

In Memory/Search

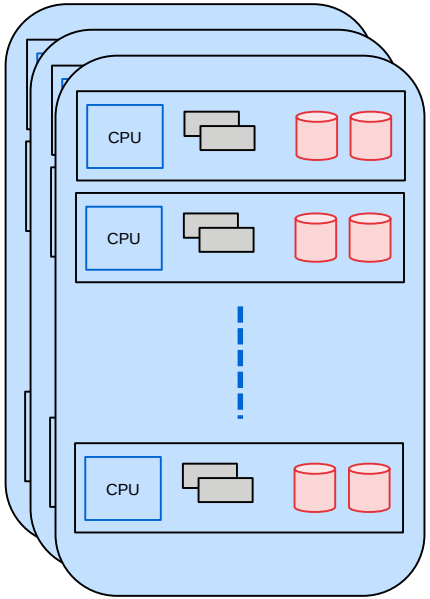
Databases

Hadoop

Object Store
FE/Dev

Archival/Cold

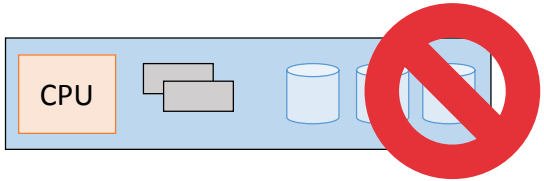
Typical hyper scale servers: Design goals



Efficiency:
Utilization, commonality

Growth:
Performance, Capacity

Converged Infrastructure: challenges



Mismatched App Needs:

- Compute/Storage needs can change for different clusters.
- Can result in under-performance or wastage

Inefficiency:

- Spend, infrastructure space, power utilization
- Further challenge to justify high density/high performance drives

Scale Challenge:

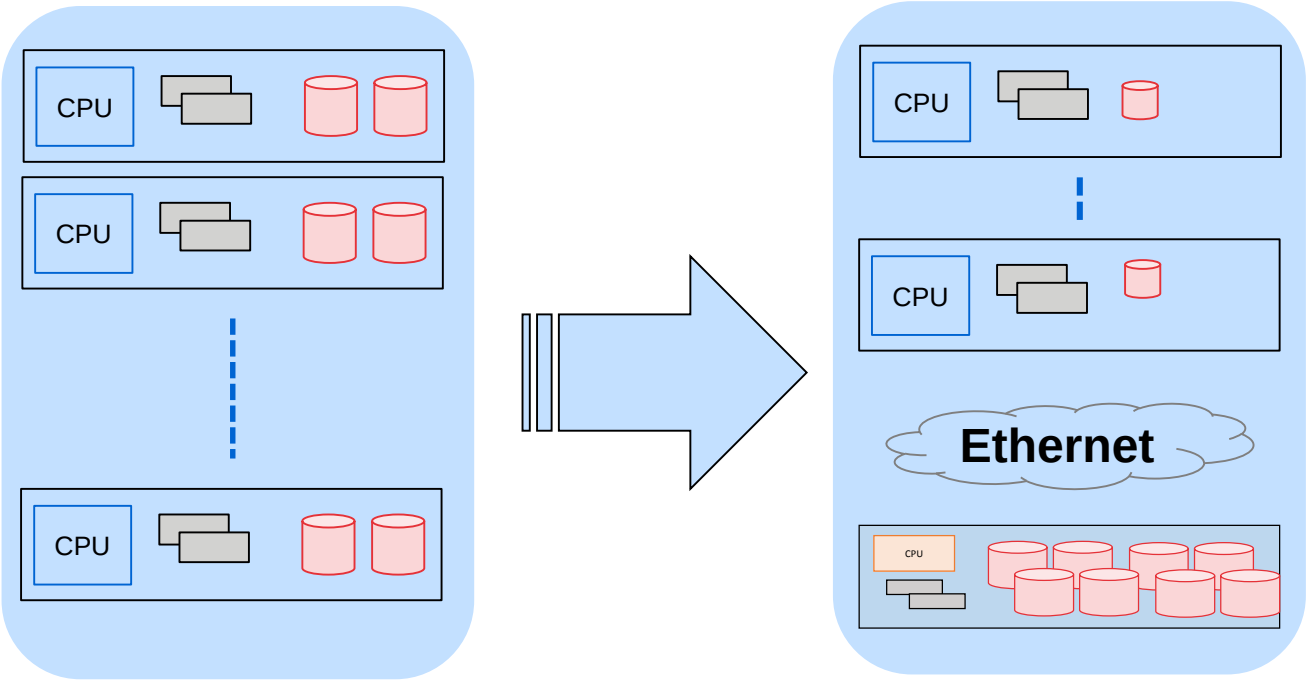
- “Cattle” use cases for data-heavy workloads may result in large data movement
- Complicated storage scheduler leads to constrained scaling

Server Platform: Shackled to local storage

What's needed: Disaggregated Storage

Separate out storage and compute resources

What's needed: Rack-As-A-Compute



Node Local Storage

Rack Local Storage

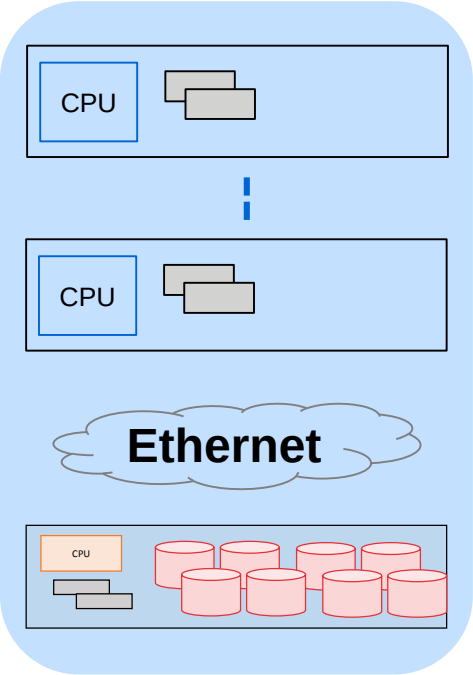
Why: Rack-As-A-Compute

Right Sizing:

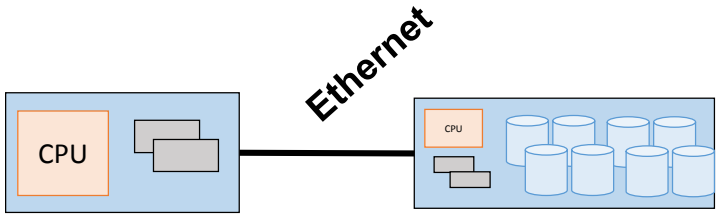
- Clusters can use optimized ratio of compute and storage.
- Allows reducing wastage and improve performance

Independent Scaling:

Compute and storage capacities can be scaled per need



Disaggregated Storage: Requirements



Throughput and IOPs:

- Sequential workloads driven by throughput
- Aggregated storage drives higher needs

IOPs:

- Smaller IO's
- Read dominated

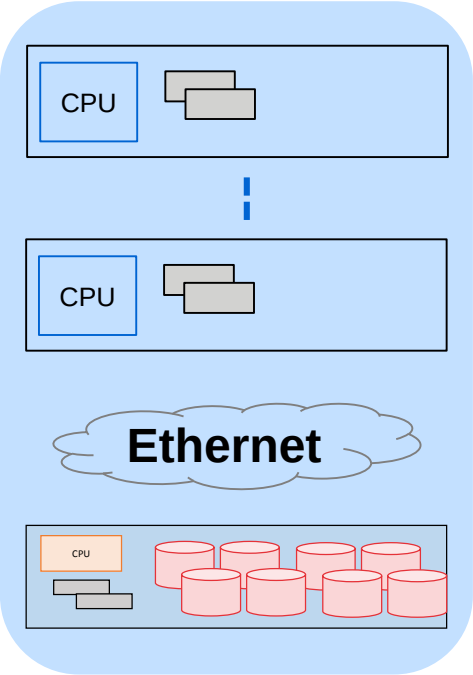
Latency:

- IOPs sensitive workloads
- Appropriate deployment topologies

Ethernet – ubiquitous, simple

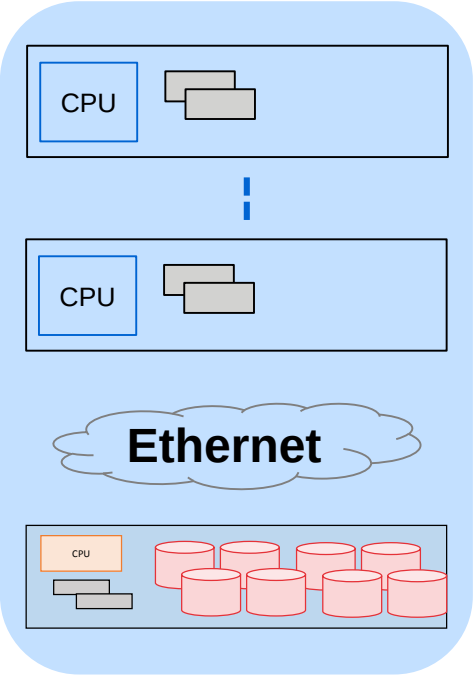
Performance Example: Disaggregated Storage

- **Storage : Compute Ratio**
  - Storage Node capacity vs. Compute Node needs
  - SN performance vs. CN needs
    - Delivered over the network
- **Example**
  - CN with 2x2TB capacity and performance need
  - Requires SN with 10x4TB NVMe
    - 10 CNs
    - ~10GBps (80 Gbps), ~1M IOPs
    - Latency : near local - average and 95th percentile



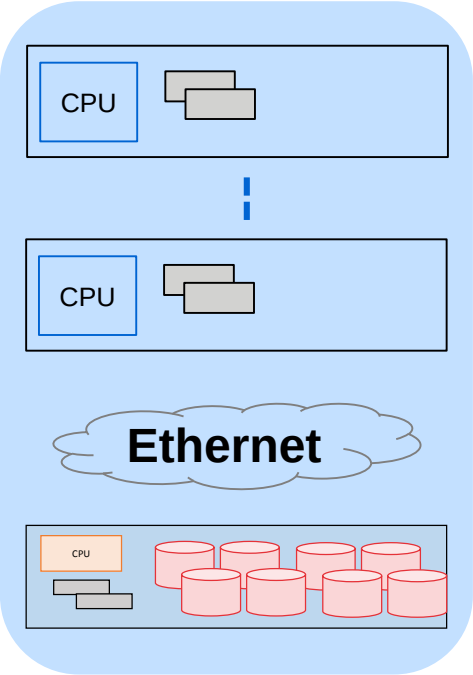
Storage Stack: Disaggregated Storage

- **Wire protocols (Block):**
  - iSCSI
  - NVMeOF : RDMA
  - NVMeOF: TCP
- **Storage Node**
  - Kernel vs. User space
    - Storage drivers
    - Storage stack
    - Network stack



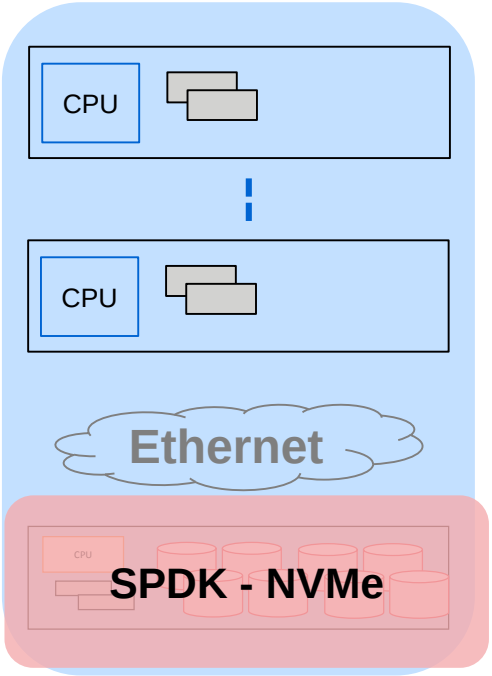
SPDK: NVMeOF thoughts:

- **RDMA**
  - Topology: Within Rack vs. Across Racks
    - RoCE, iWARP
- **TCP**
  - Good enough performance?
  - SPDK support?



Disaggregated Storage

Are we there yet?



Thank You

ebay