SDC
STORAGE DEVELOPER CONFERENCE
SNIA ■ SANTA CLARA, 2014

Building Cost-Effective, Scale-Out All-Flash Systems

Doron Tal, Chief Architect

kaminario.

Agenda

- What is “cost-effective” on AFA?
- Scale-up and scale-out
- Real-life capacity and performance metrics
- Design principals
  - Efficient RAID
  - Scalable metadata
  - Adaptive block size
  - Scale-out

What Do We Mean When We Say "Cost-Effective"?

- Performance $/[IO,Throughput]
- Capacity $/raw GB, $/effective GB
- Energy Watt/GB, Watt/IO
- GB/RU, performance/RU

Flexibility to meet different requirements

All-Flash Array (AFA) Considerations

- Performance is mainly governed by CPU
  - 24 SSDs can theoretically provide over 2M IOPS
  - Similar performance will require thousands of HDDs
- Capacity requires significant resources
  - Granular thin provisioning, deduplication and compression requires significant amounts of resources for metadata handling and caching

Scale-Up vs. Scale-Out

Scale-Up

- Performance $/[IO,Throughput]
- Capacity $/GB
- Energy Watt/GB, Watt/IO
- GB/RU, performance/RU

Flexibility to meet different requirements

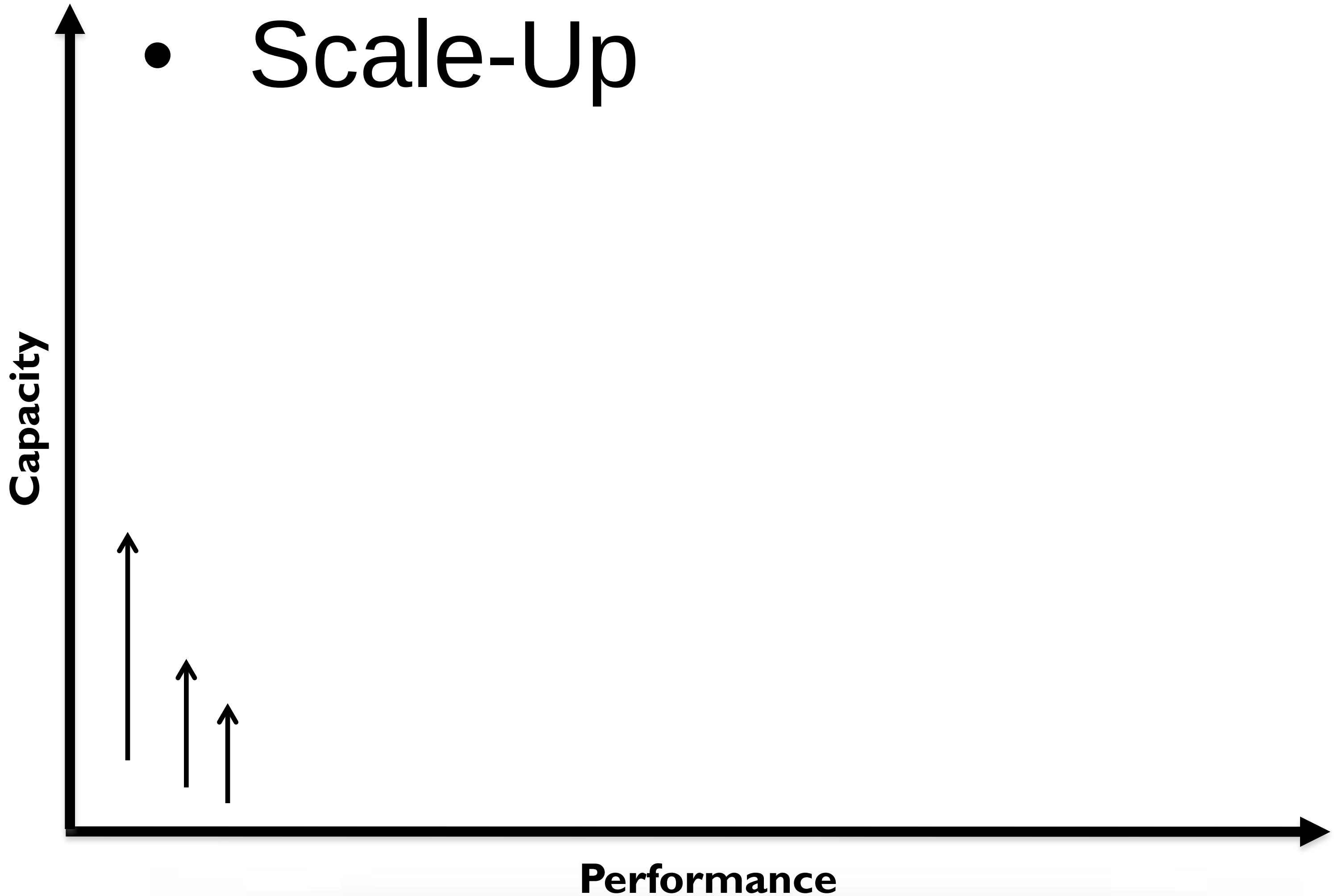
Scale-Out

- Performance $/[IO,Throughput]
- Capacity $/GB
- Energy Watt/GB, Watt/IO
- GB/RU, performance/RU

Flexibility to meet different requirements

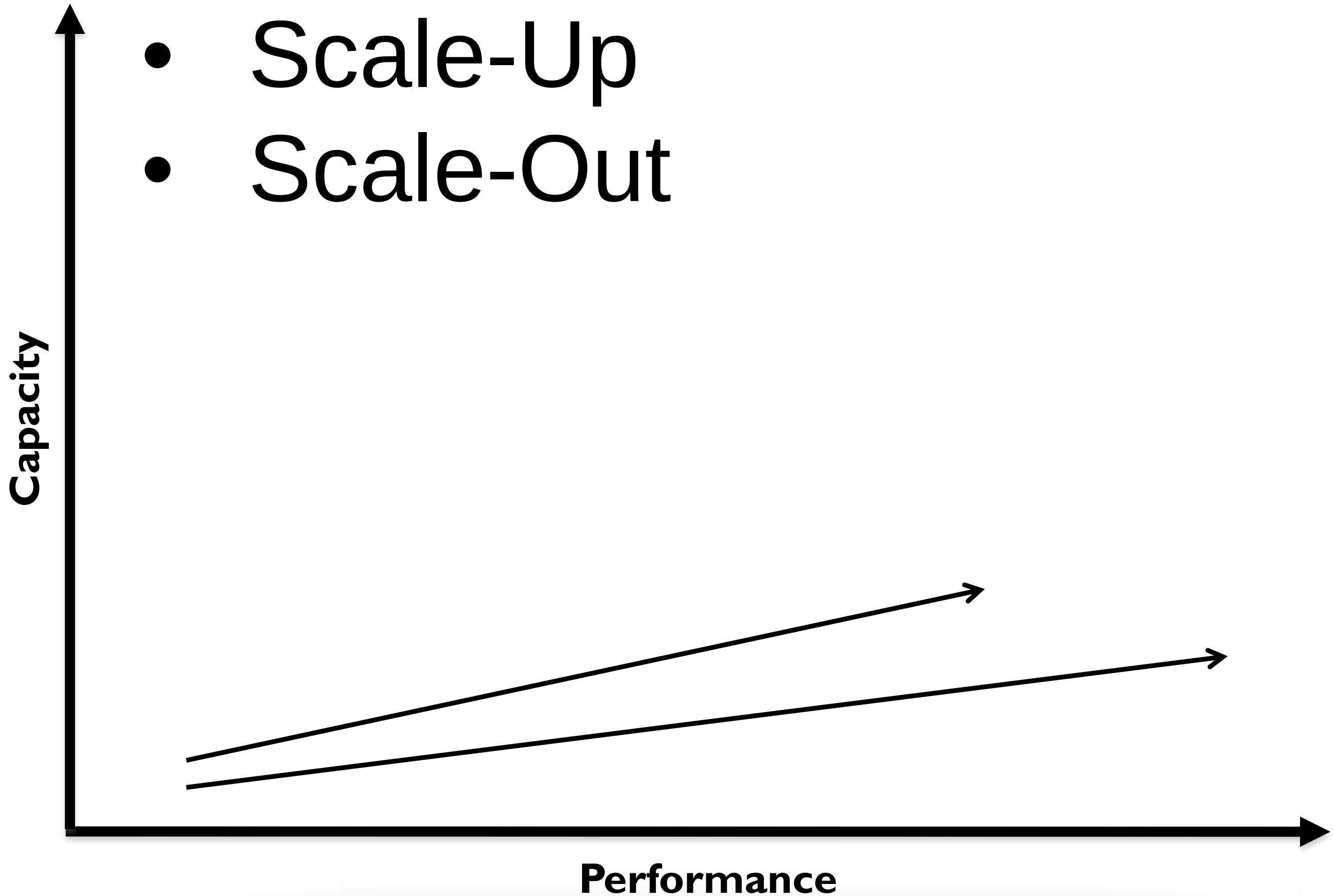
AFA Capacity Performance Graph

- Scale-Up



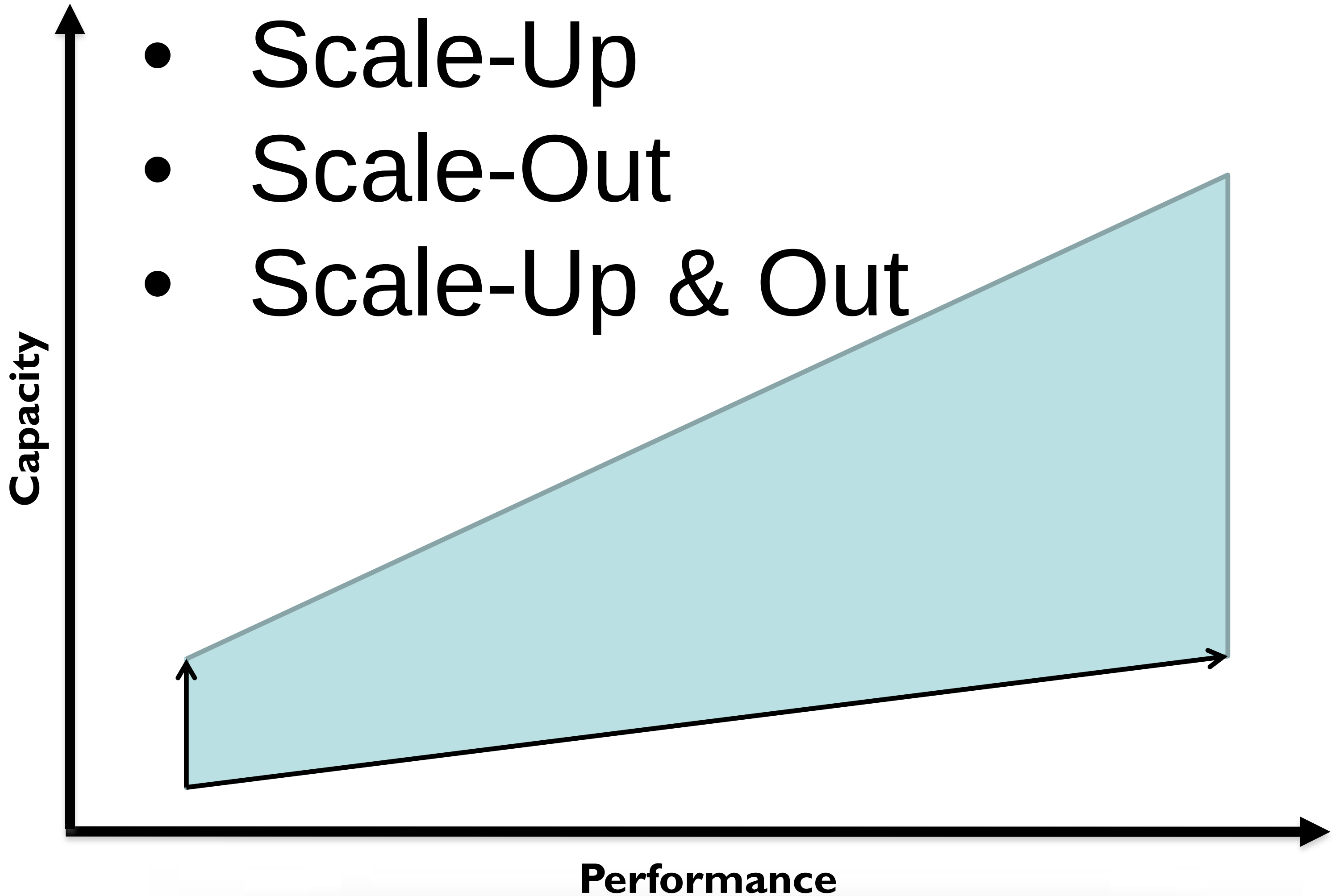
Capacity Performance

- Scale-Up
- Scale-Out

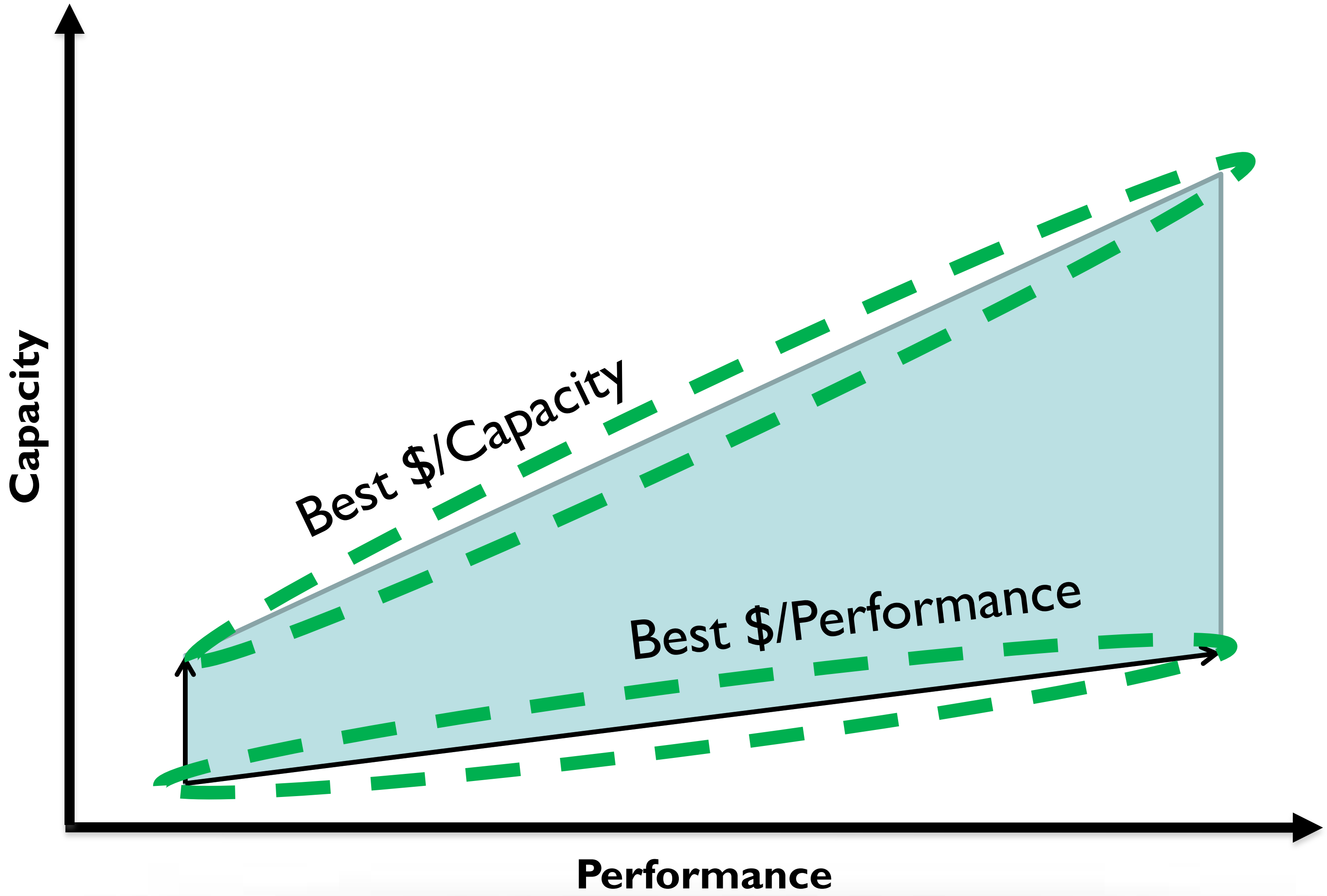


Capacity Performance

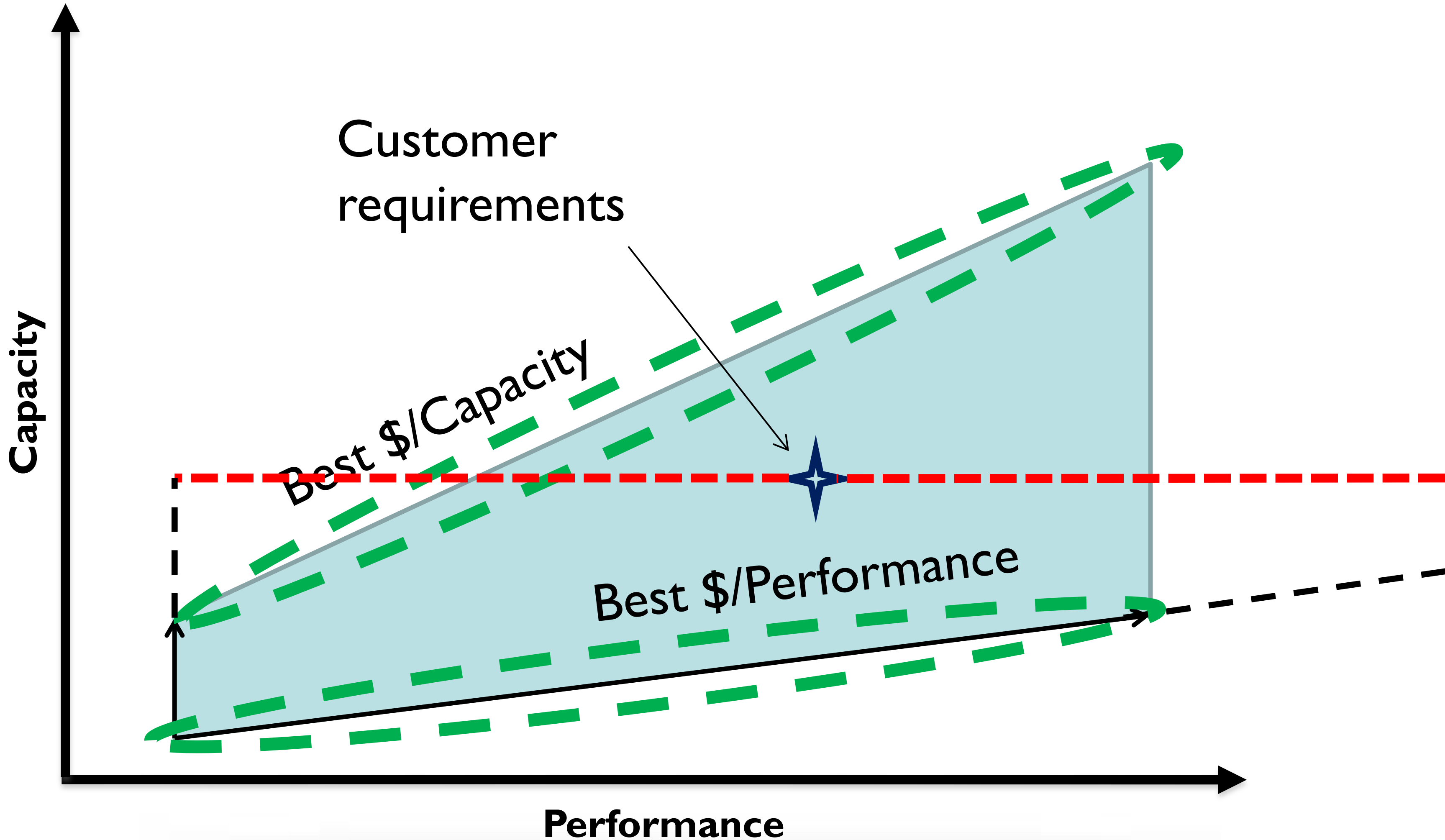
- Scale-Up
- Scale-Out
- Scale-Up & Out



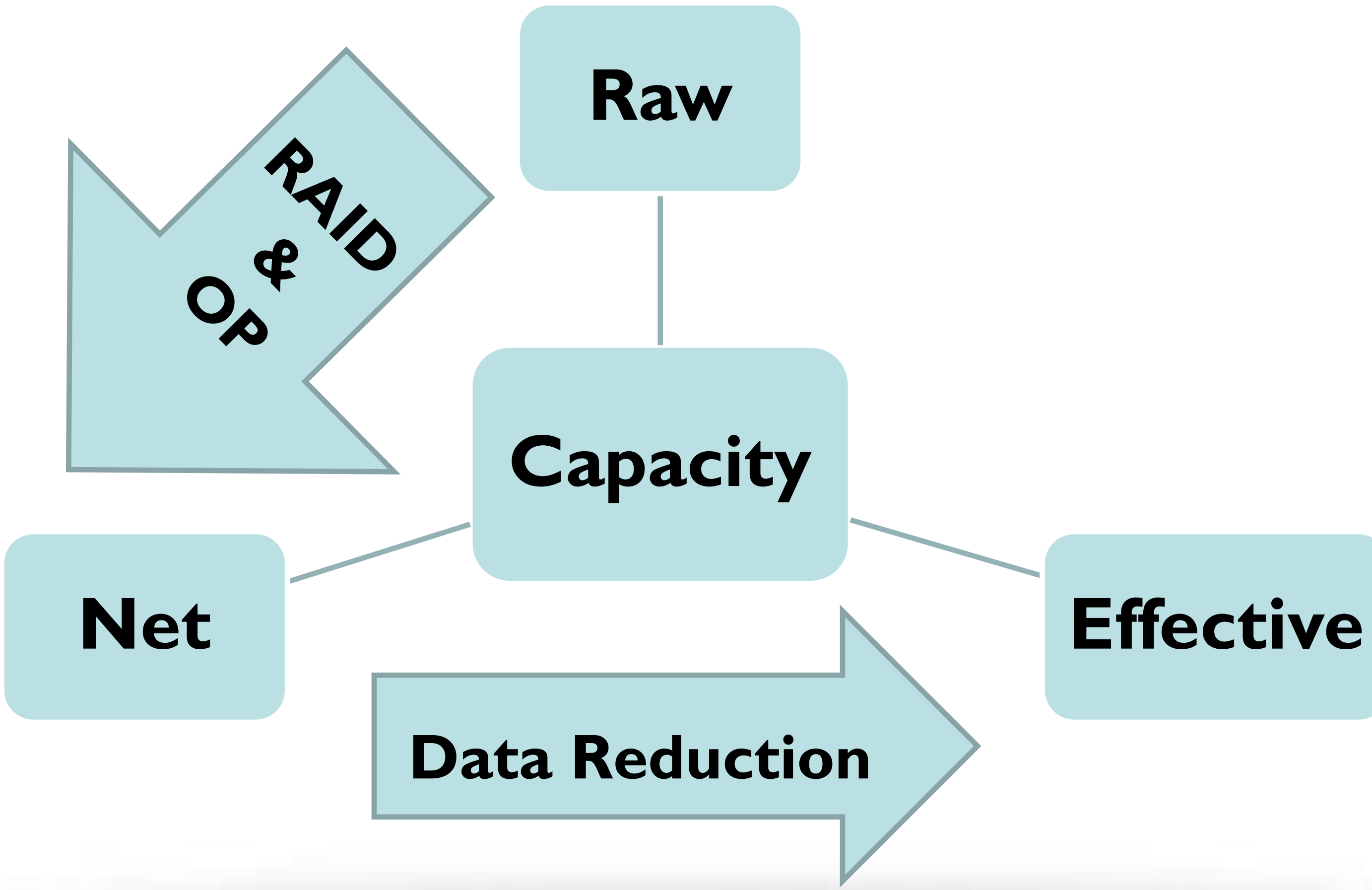
Capacity Performance



Capacity Performance



Capacity Metrics

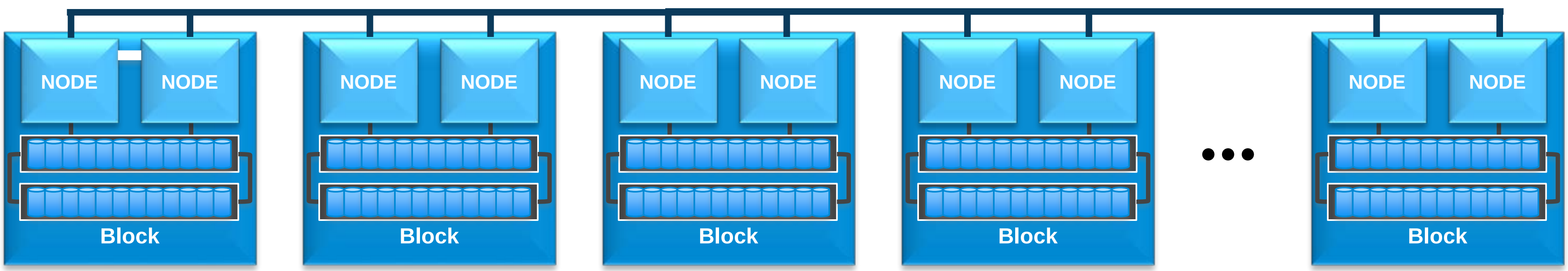


Performance Metrics

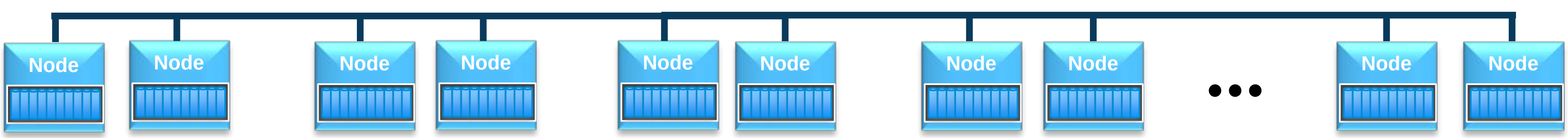
- Real world vs. synthetic benchmark
- Synthetic benchmark examples
  - Maximum 4KB read IOPS
  - Sequential write
  - Small set of duplicated data
  - More

RAID

- Two popular approaches
  - Shared storage RAID



  - Direct Attached Storage (DAS) RAID



RAID - Efficiency

Efficiency after protection and spare for recovery

- RAID efficiency

Nodes #	2	4	8	16
Shared RAID 6	21/24 =87.5%	21/24 =87.5%	21/24 =87.5%	21/24 =87.5%
DAS mirror		0.5*(3/4) =37.5%	0.5*(7/8) =~44%	0.5*(15/16) =~47%
DAS RAID 6		1/4 =25%	5/8 =62.5%	13/16 =81.25%

- Main advantages of shared RAID 6 architecture
  - Real entry-level option
  - Higher efficiency on medium to large clusters

RAID – Failure Domain and its impact on scale-up flexibility

kaminario

- Largest failure domain that requires data movement depends on architecture
- Shared – failure domain is SSD
  - More capacity per server (more SSDs) -> same failure domain
- DAS – failure domain is Server
  - More capacity per server -> more data to move
  - DAS advantage: More servers -> faster recovery
  - More capacity per server X more servers = **<u>Lots of capacity</u>** (more than required!)

Scalable Metadata

- Pointer per 4KB of addressable data – LU/LBA to address type 1
- Pointer per 4KB of unique data - Content signature to address type 2
- Interesting properties:
  - #Type 1 / #Type 2 = Deduplication ratio
  - #Type 2 / Net capacity = Compression ratio
- Sizeof(type 1) ~= 8Bytes
- Sizeof(type 2) ~= 8Bytes (weak hash)
  20Bytes (strong hash)

Scalable Metadata

- Compression 2:1, Deduplication 3:1

Net	1TB			10TB			100TB		
	Type 1	Type 2	Total	Type 1	Type 2	Total	Type 1	Type 2	Total
Weak Hash	12GB	4GB	32GB	120GB	40GB	320GB	1.2TB	400GB	3.2TB
Strong Hash	12GB	10GB	44GB	120GB	100GB	440GB	1.2TB	1TB	4.4TB

- Total column includes 2X spare for server failure
- In many cases data reduction can be > 6
- Conclusion: Not all metadata can reside in RAM

Scale-Out

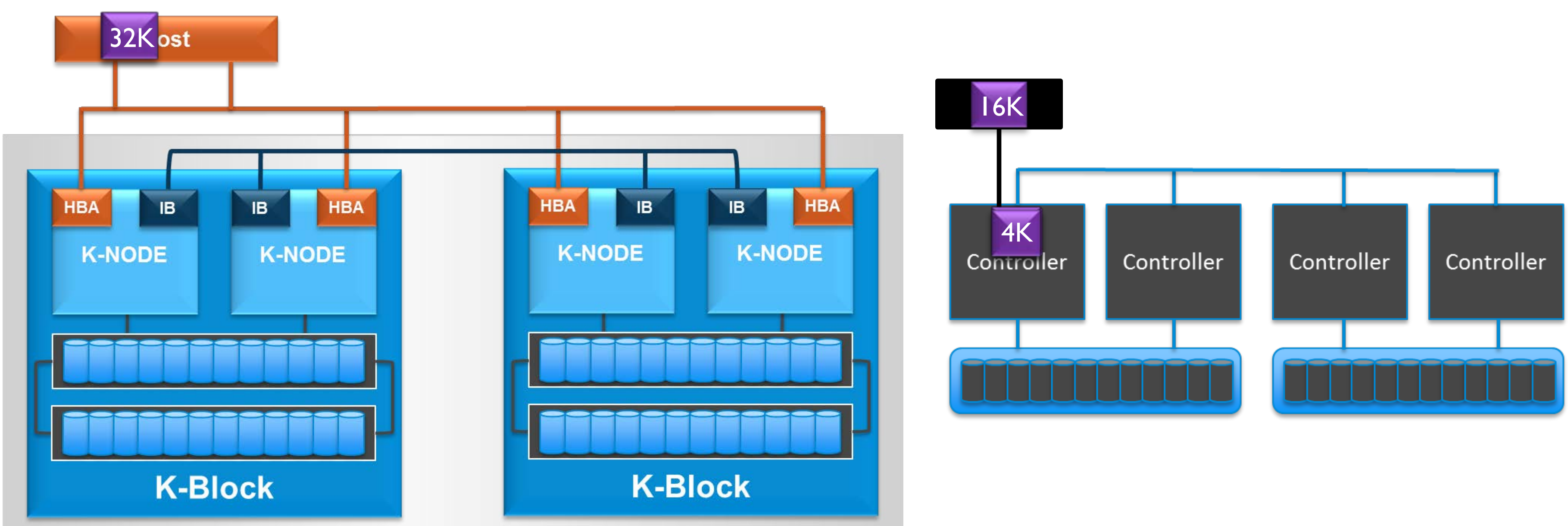
- Consistent performance
- Agility and flexibility
  - Single system
  - No forklift upgrades
- Global deduplication
- Full Active/Active
- Best $/performance

Adaptive Block Size

- Eliminating any 4KB match globally in the system
- Break IO to the largest consecutive match found

Adaptive block size architecture

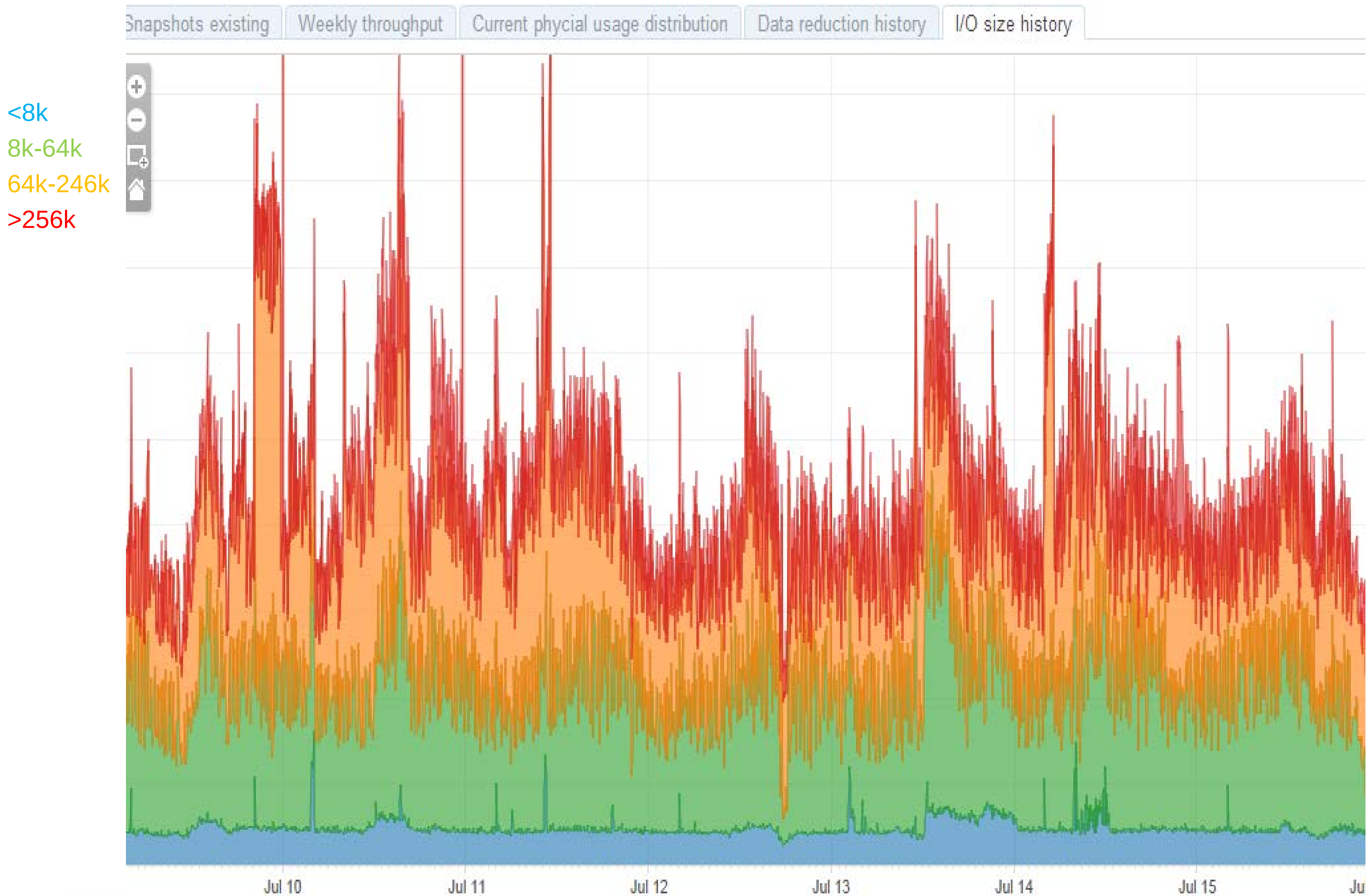
Fixed 4KB content address mapping architecture



A read is processed as a **SINGLE** IO operation

A read is processed as **MULTIPLE** 4KB blocks kaminario.

Field Block Size Statistics

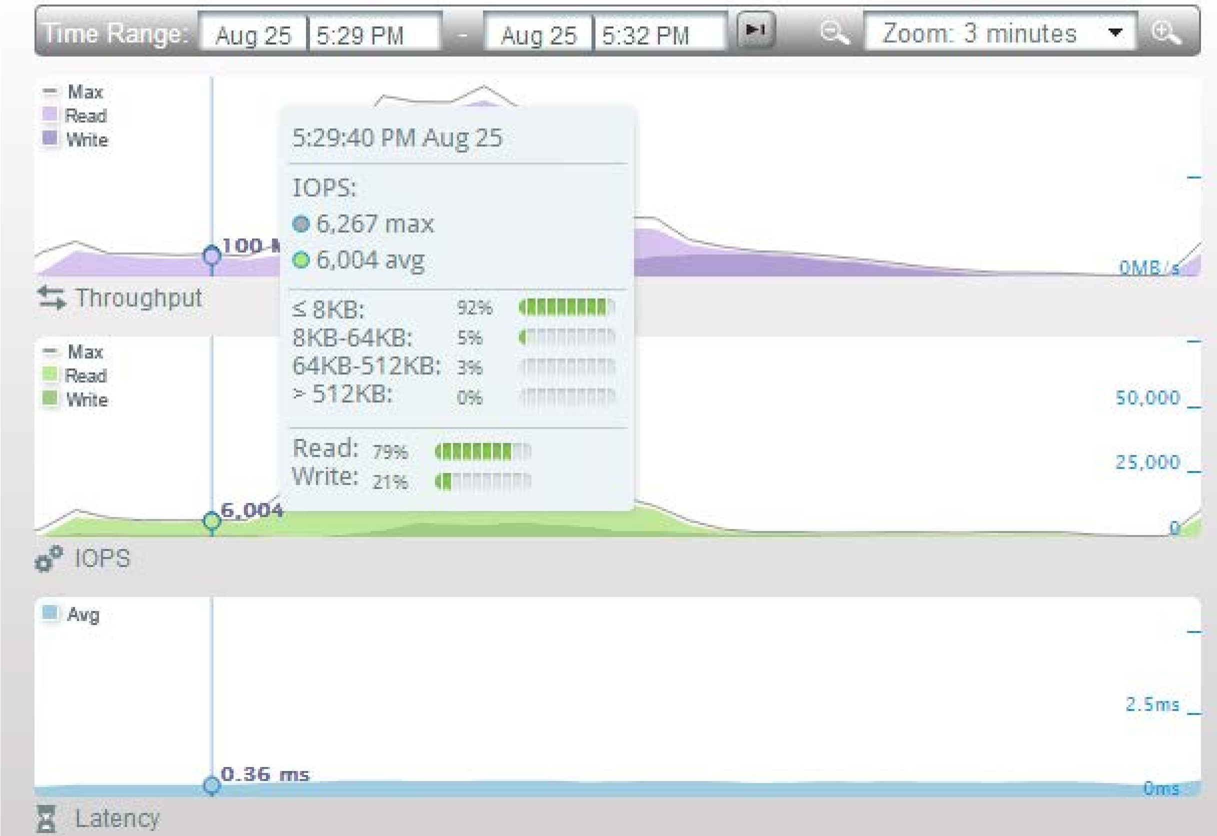


SDC 14

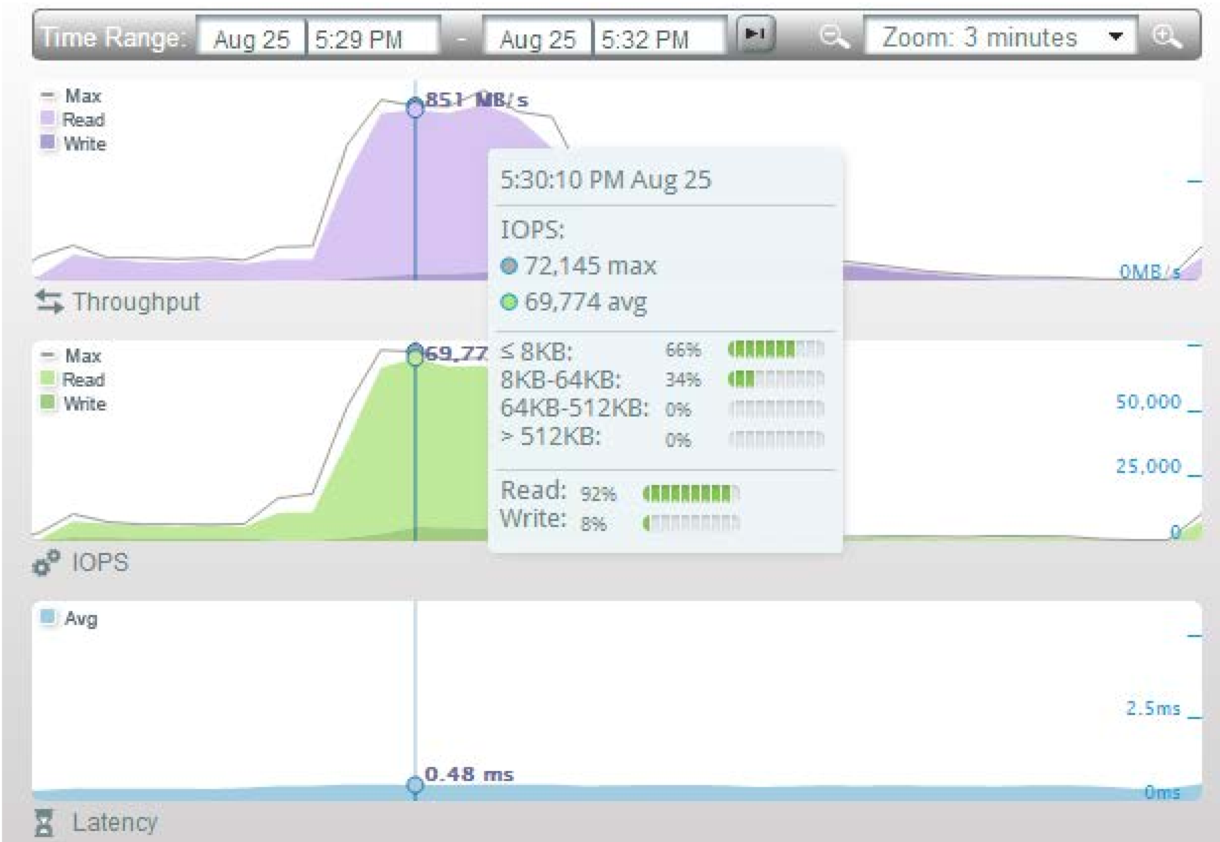
Common I/O Sizes Per Applications

- Databases
  - OLTP: 8KB-64KB random
  - OLAP: & DWH > 64KB
  - LOG: Sequential
  - TempDB: similar to OLTP
- Exchange: 32KB random
- VDI
  - Clone: sequential large blocks (XCOPY)
  - Boot storm: is it 4KB?

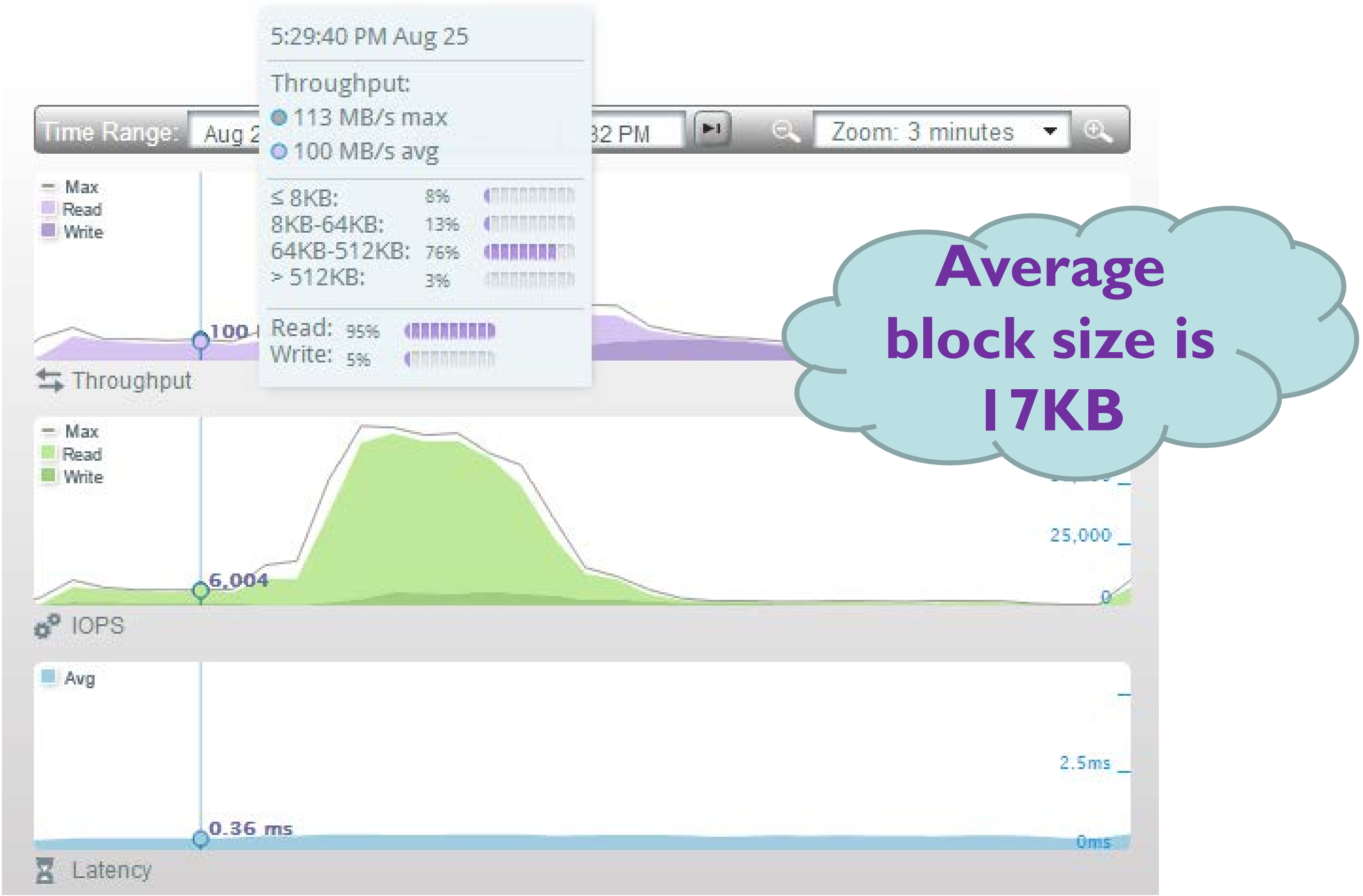
VDI Boot Storm



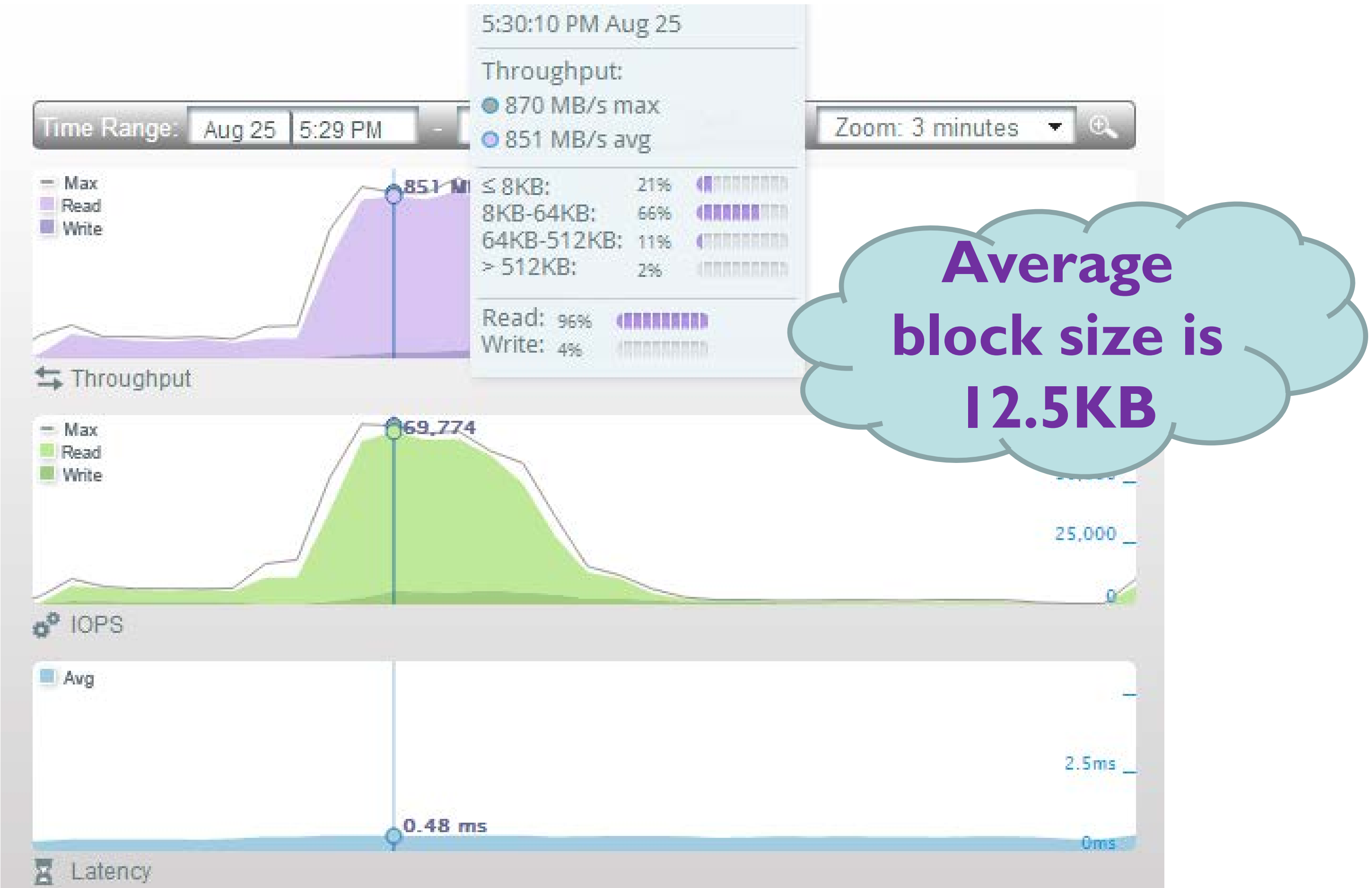
VDI Boot Storm



VDI Boot Storm



VDI Boot Storm



Adaptive Block Size Advantages

- CPU utilization
  - Less I/O
  - Less messages
- Metadata
  - Denser metadata
  - Less metadata retrieval

Questions?

doron.tal@kaminario.com
kaminario.com