SDC
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2015

# Pelican: A Building Block for Exascale Cold Data Storage

**Austin Donnelly**
**Microsoft Research**

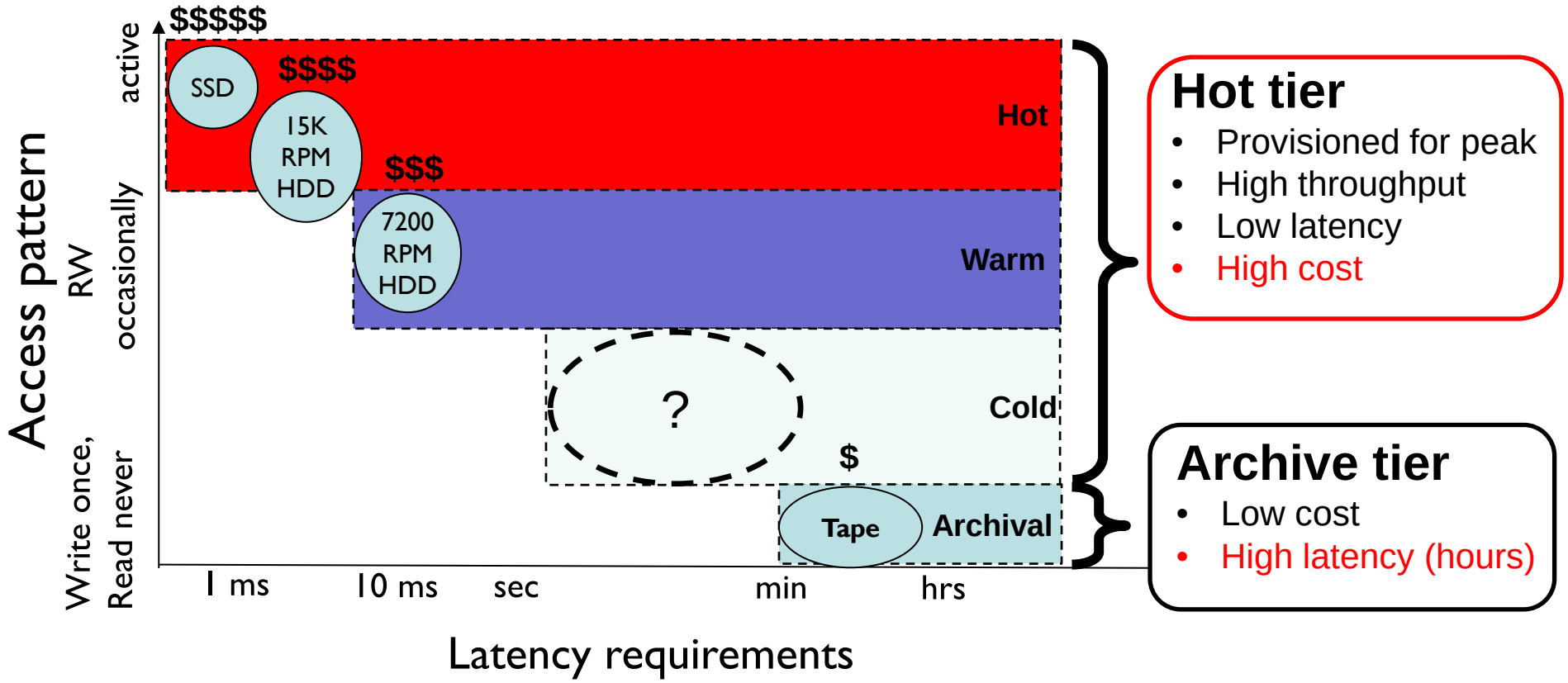
and: Shobana Balakrishnan, Richard Black, Adam Glass, Dave Harper, Sergey Legtchenko, Aaron Ogus, Eric Peterson, Ant Rowstron

# Outline

- Background
- Pelican co-design
- Research challenges
- Demo
- Performance results


2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

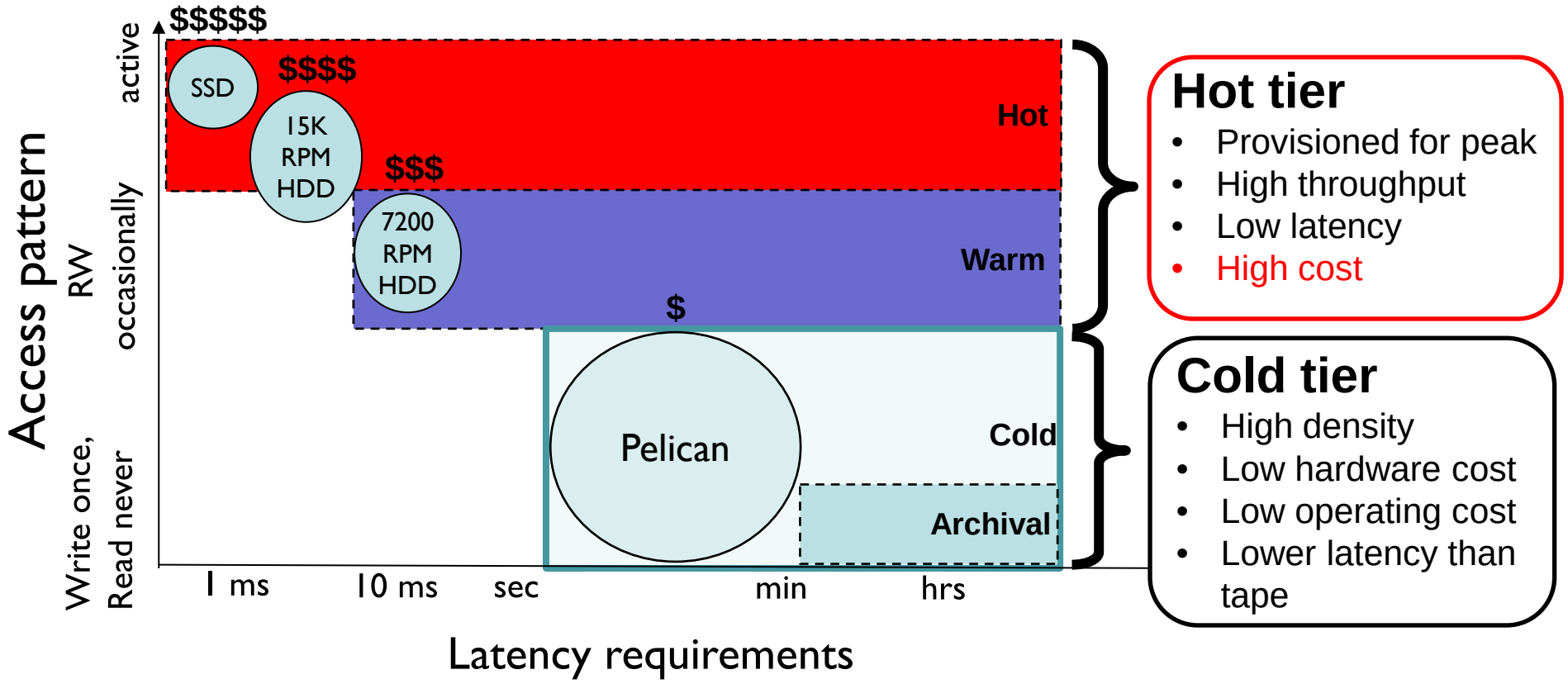
# Background: Cold Data in the Cloud






2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# Background: Cold Data in the Cloud




## Hot tier

- Provisioned for peak
- High throughput
- Low latency
- High cost

## Cold tier

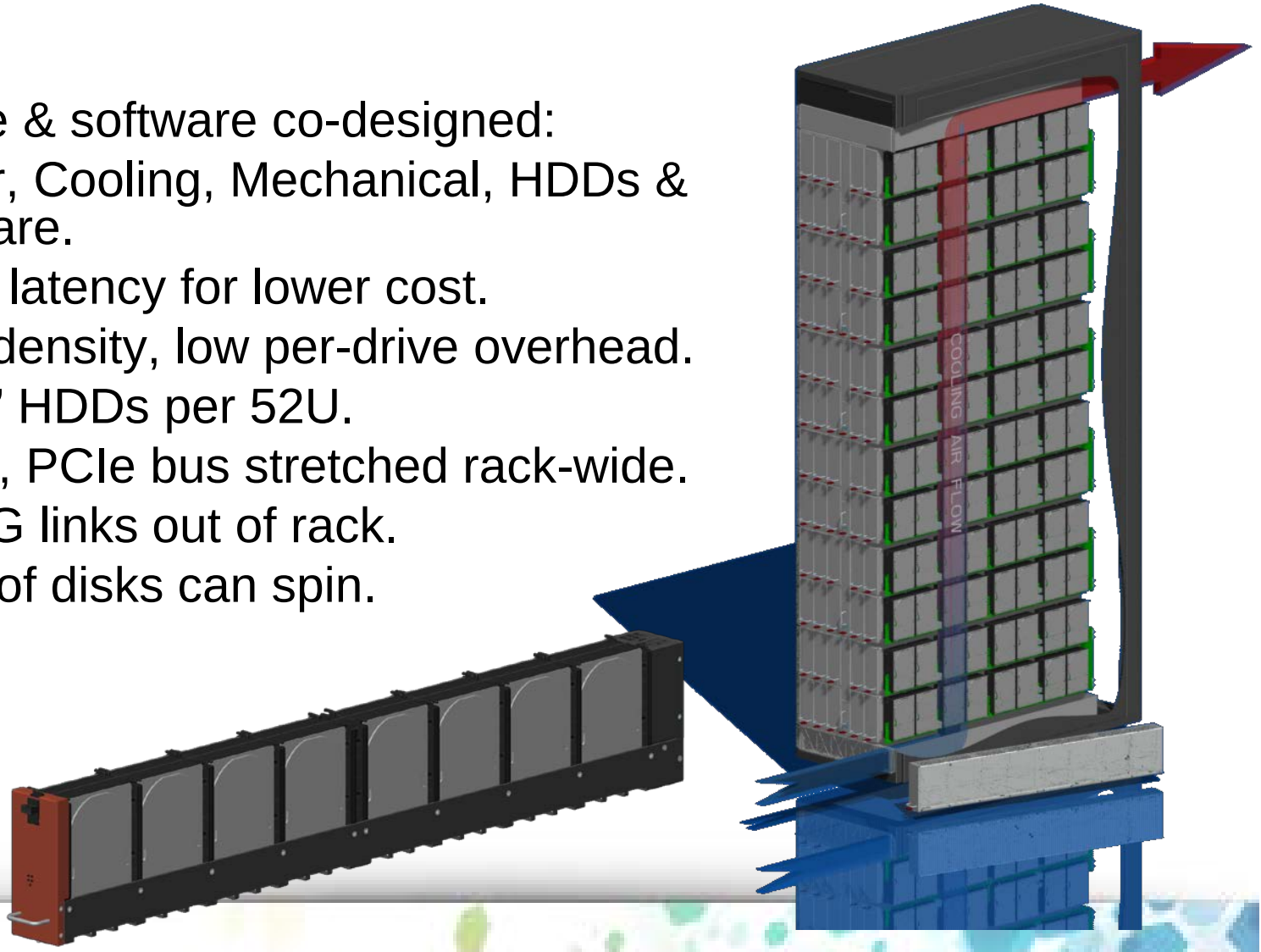
- High density
- Low hardware cost
- Low operating cost
- Lower latency than tape



2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

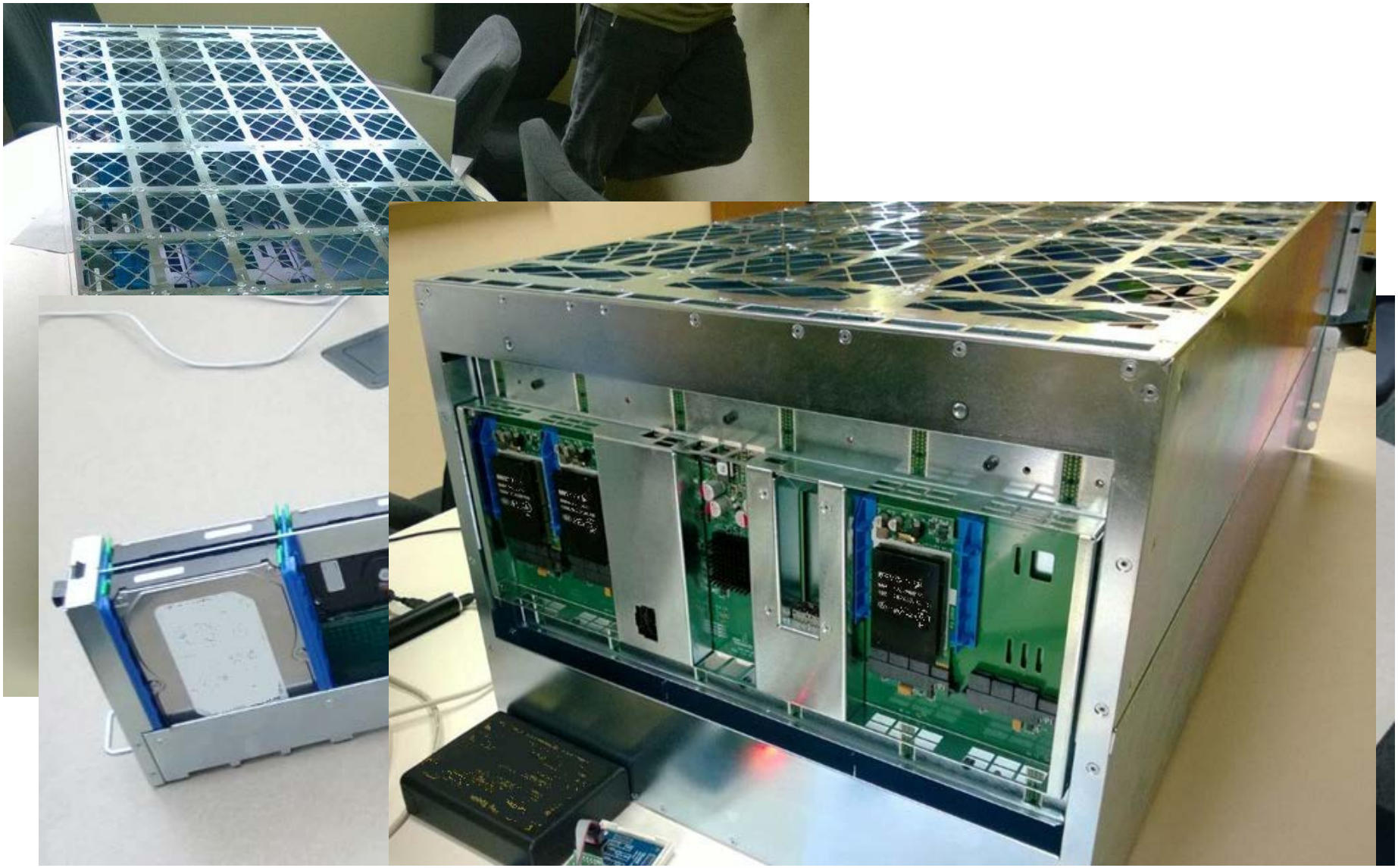
# Pelican: Rack-scale Co-design

- Hardware & software co-designed:
  - Power, Cooling, Mechanical, HDDs & Software.
  - Trade latency for lower cost.
- Massive density, low per-drive overhead.
- 1152 3.5” HDDs per 52U.
- 2 servers, PCIe bus stretched rack-wide.
  - 4x 10G links out of rack.
- Only 8% of disks can spin.





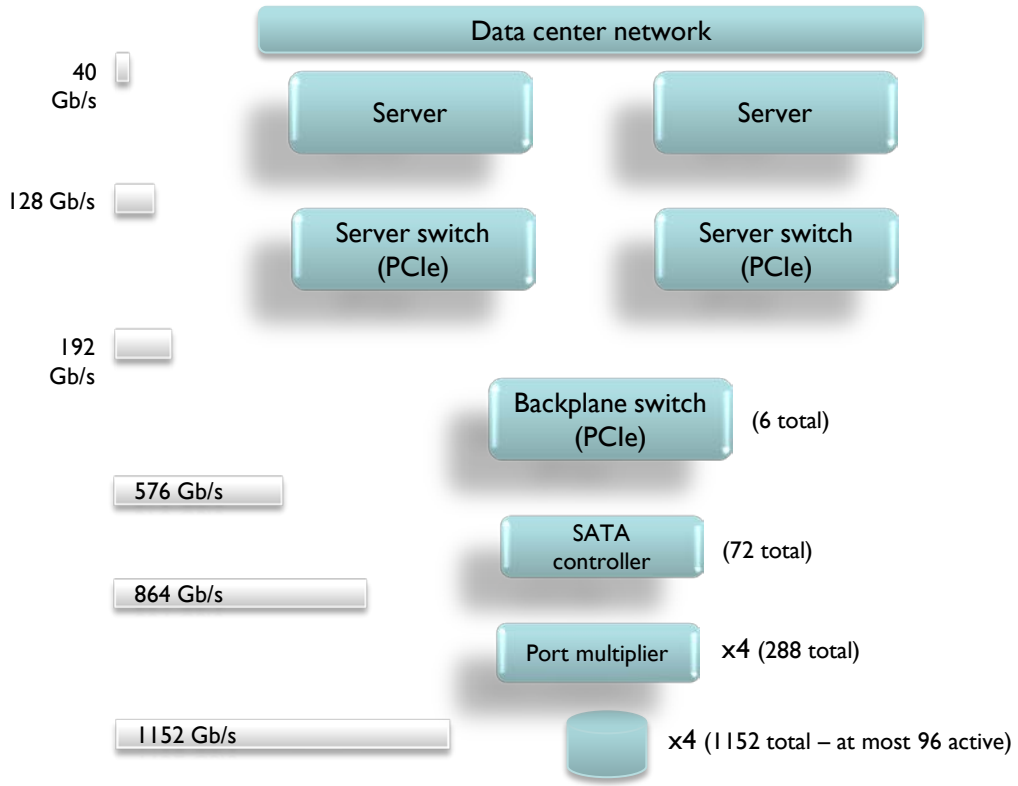

2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.





2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# Interconnect Details




Bandwidth required:

96 Gb/s

96 Gb/s

96 Gb/s

96 Gb/s

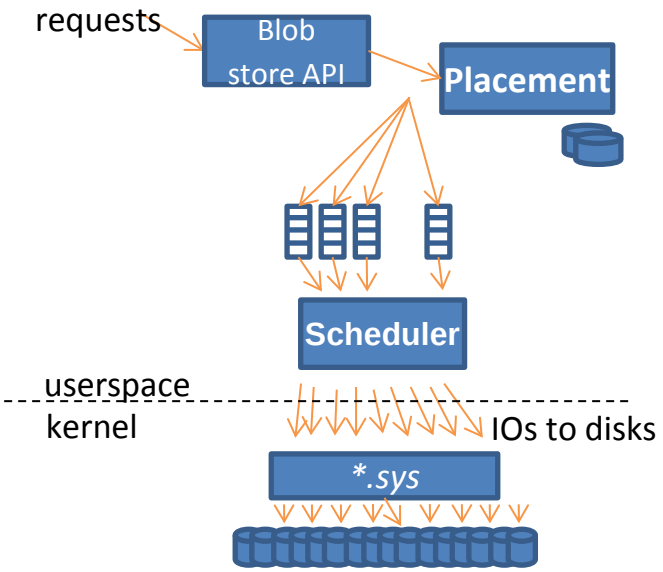


2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# Research Challenges

- Not enough cooling, power, or bandwidth.
- How do we manage theses resource limits?
- Which disks to use for data? The **data layout** problem.
- How to **schedule** requests to get good performance.

["*Pelican: A building block for exascale cold data storage*", OSDI 2014]



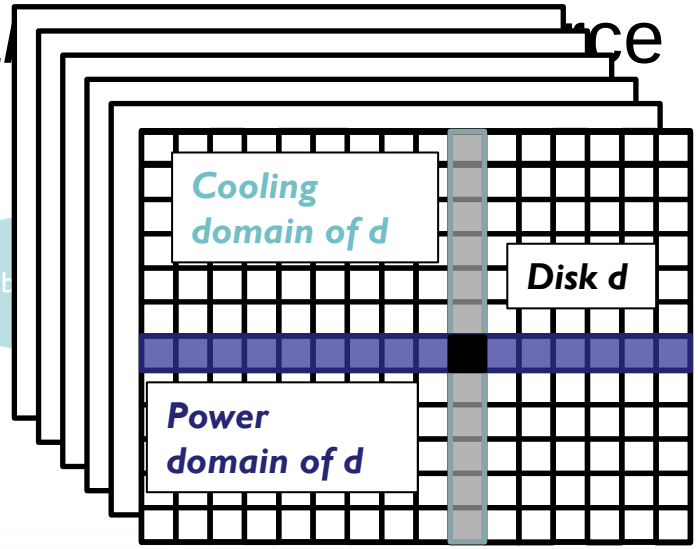


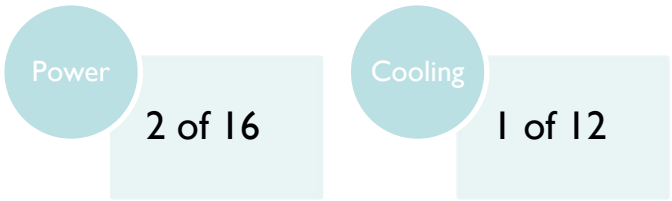
2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# Resource use

- Traditional systems:
  - Any disk can be active at any time.
- Pelican:
  - Disk is part of a *domain* of resource
- Domains, limits:

Rack: 3D array of disks






2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

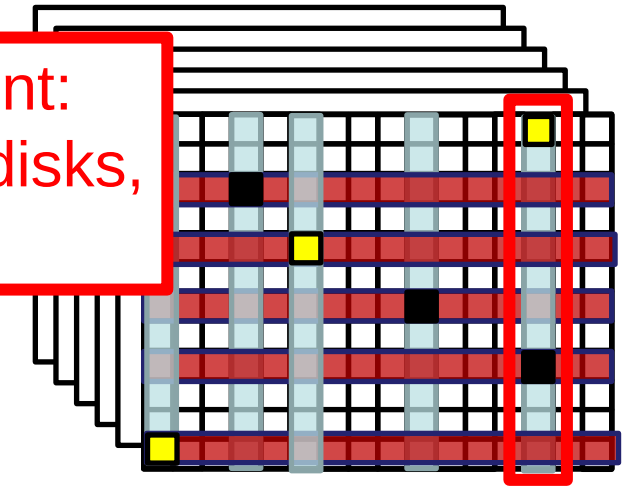
# Data placement

- Blob erasure-encoded on a ***set*** of concurrently active disks
- **In traditional systems:**
  - Any two sets can be active
  - No impact on concurrency
- **In Pelican:**
  - Sets can conflict in resource requirements
  - Conflicting sets cannot be concurrently active
- Challenge: form sets to minimize $P_{conflict}$


2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# Data placement: random

Random placement:
Storing blobs on n disks,
$P_{conflict} \rightarrow O(n^2)$



*Conflict*

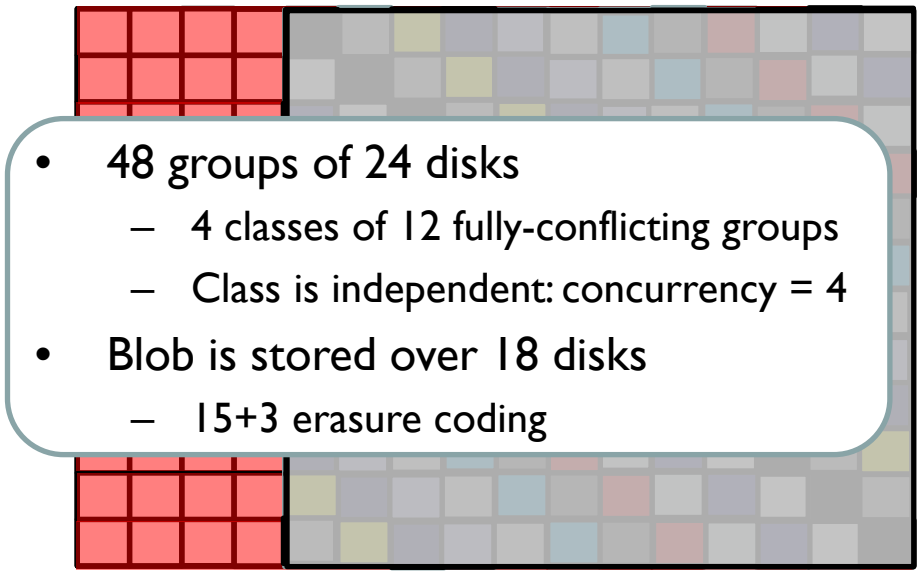
Disks of blob 1
Disks of blob 2



2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# Data placement: Pelican

- **Intuition**: concentrate conflicts over a few sets of disks.
- Store blob in one group
  - $P_{conflict}$ ➔ O(n)
- Groups encapsulate constraints:
  - Unit of IO scheduling
  - No constraints managed at runtime.




*Schematic side-view of the rack*

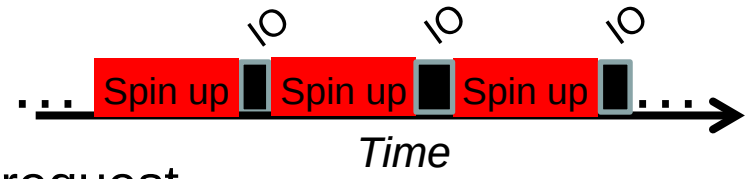

2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

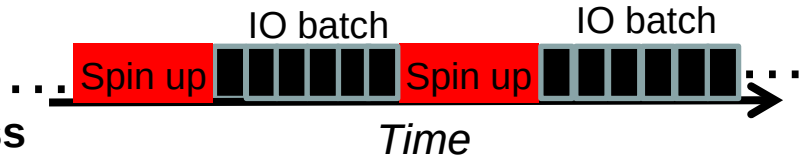
# IO Scheduling: “spin up is the new seek”

Four independent schedulers

Each scheduler: 12 groups, only one can be active

- Naïve scheduler: FIFO
  - Avg. group activation time: 14.2 sec
  - High probability of spinup after each request
  - Time is spent doing spinups!
- Pelican scheduler: Request batching
  - Limit on maximum re-ordering
  - Trade-off between **throughput** and **fairness**
  - Weighted fair-share between client and rebuild traffic





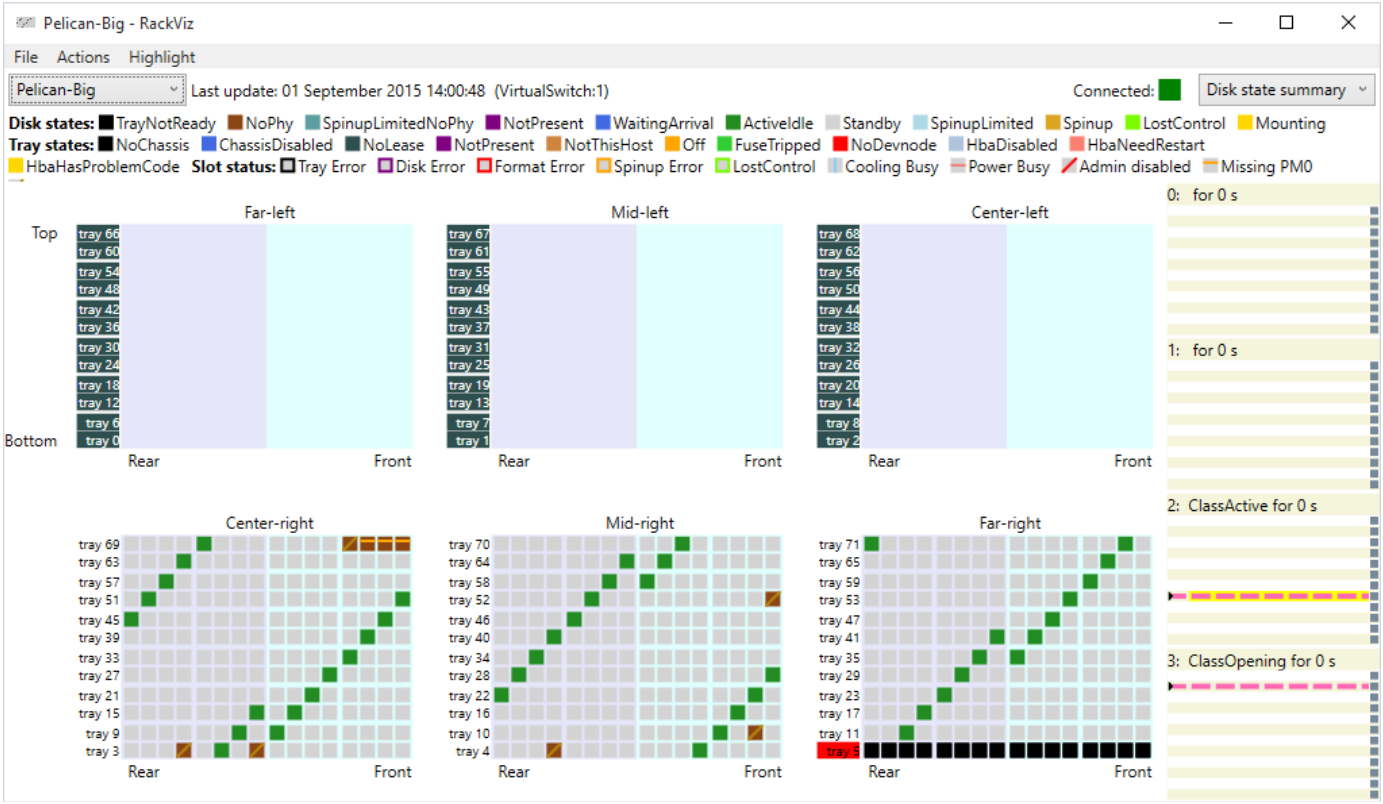

2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# Outline

- Background
- Pelican co-design
- Research challenges:
  - Data placement: constraint-aware
  - Scheduler: batching to amortize spinups
- **Demo**
- Performance results


2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# Demo




2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

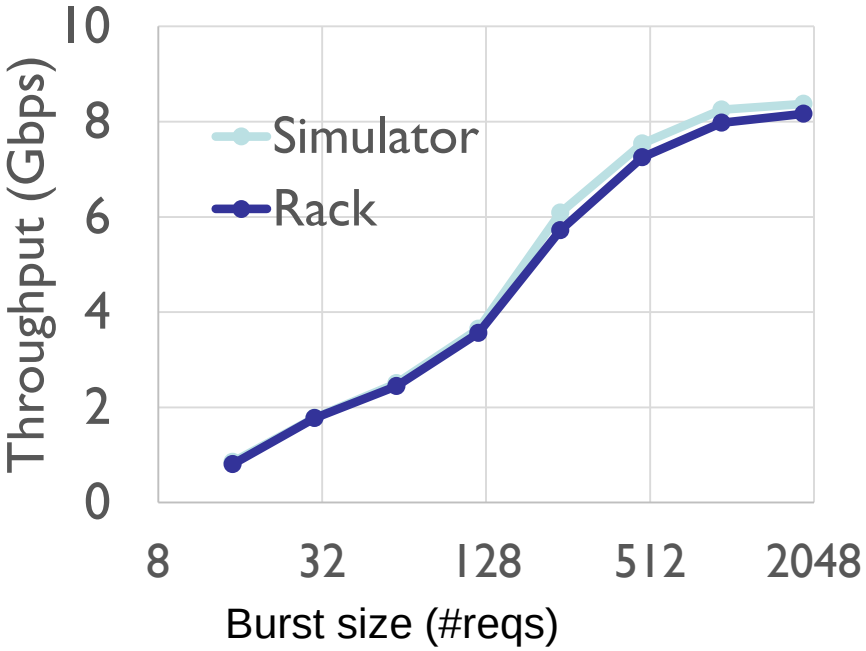
# Performance

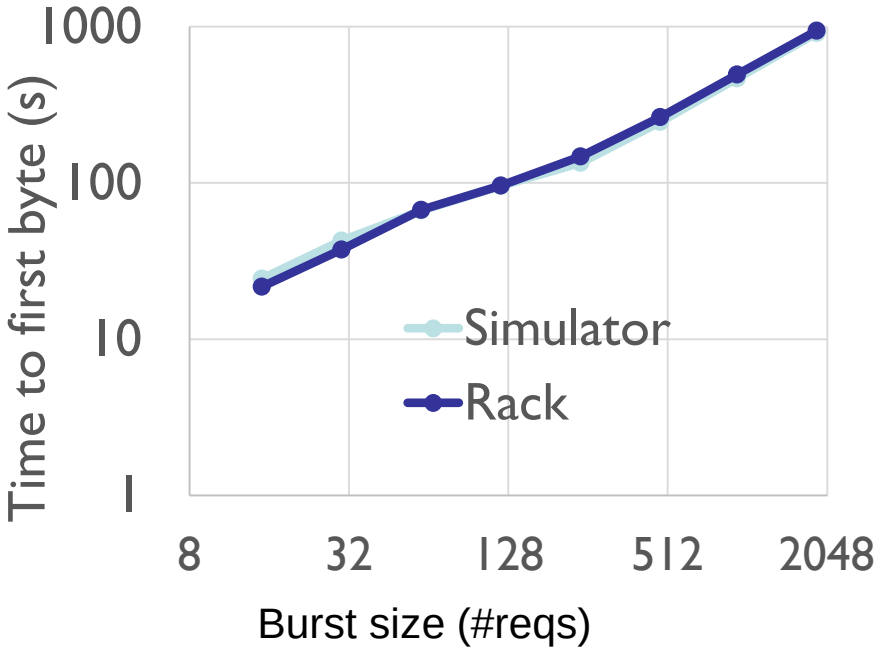
- Compare Pelican vs. all disks active (FP).
- Cross-validate simulator.
- Metrics:
  - Throughput
  - Latency (time to first byte)
  - Power consumption
- Open loop workload:
  - Poisson arrivals
  - Read requests, 1GB blobs


2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# First step: simulator cross-validation

- Burst workload, varying burst size

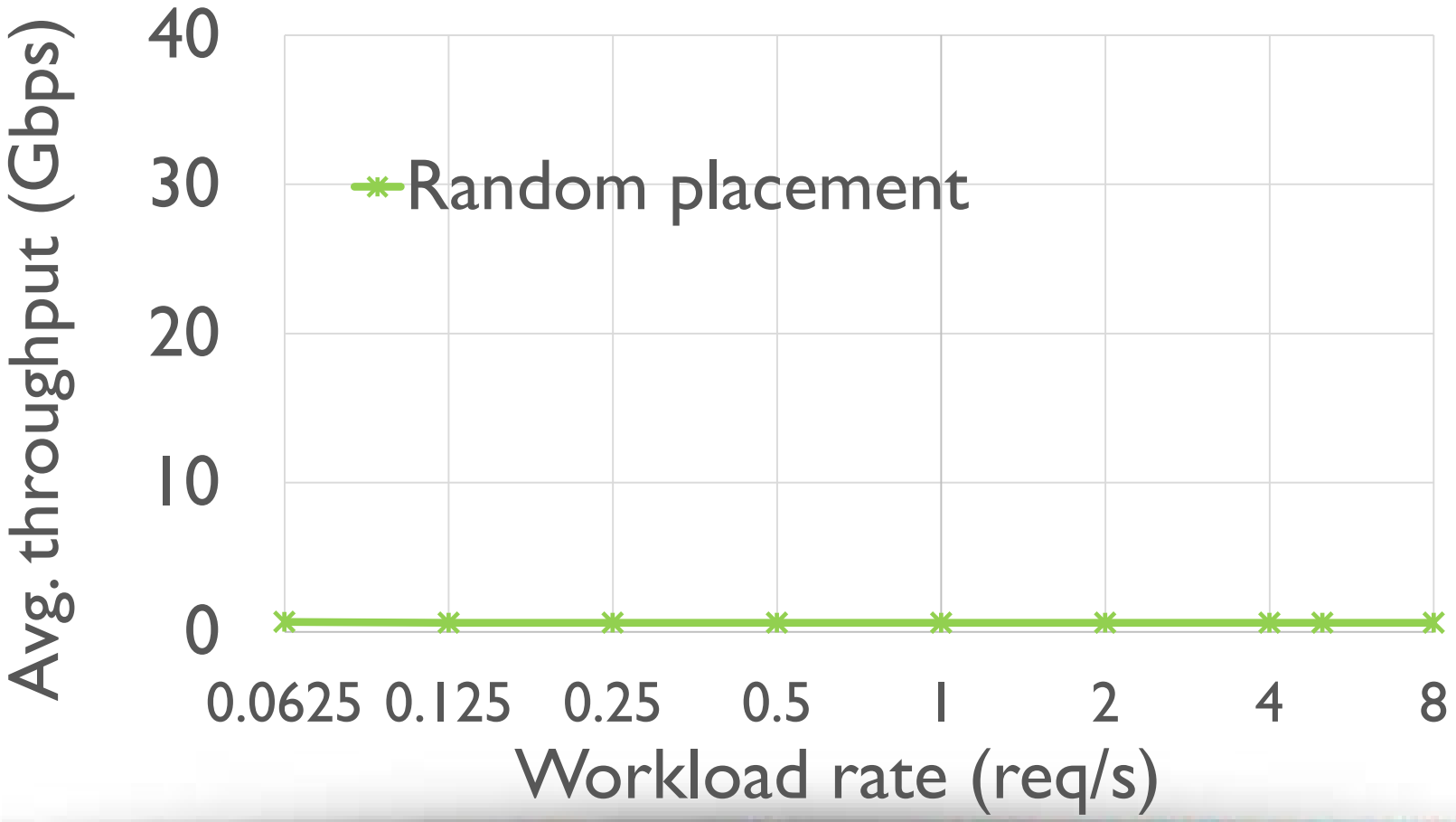





2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

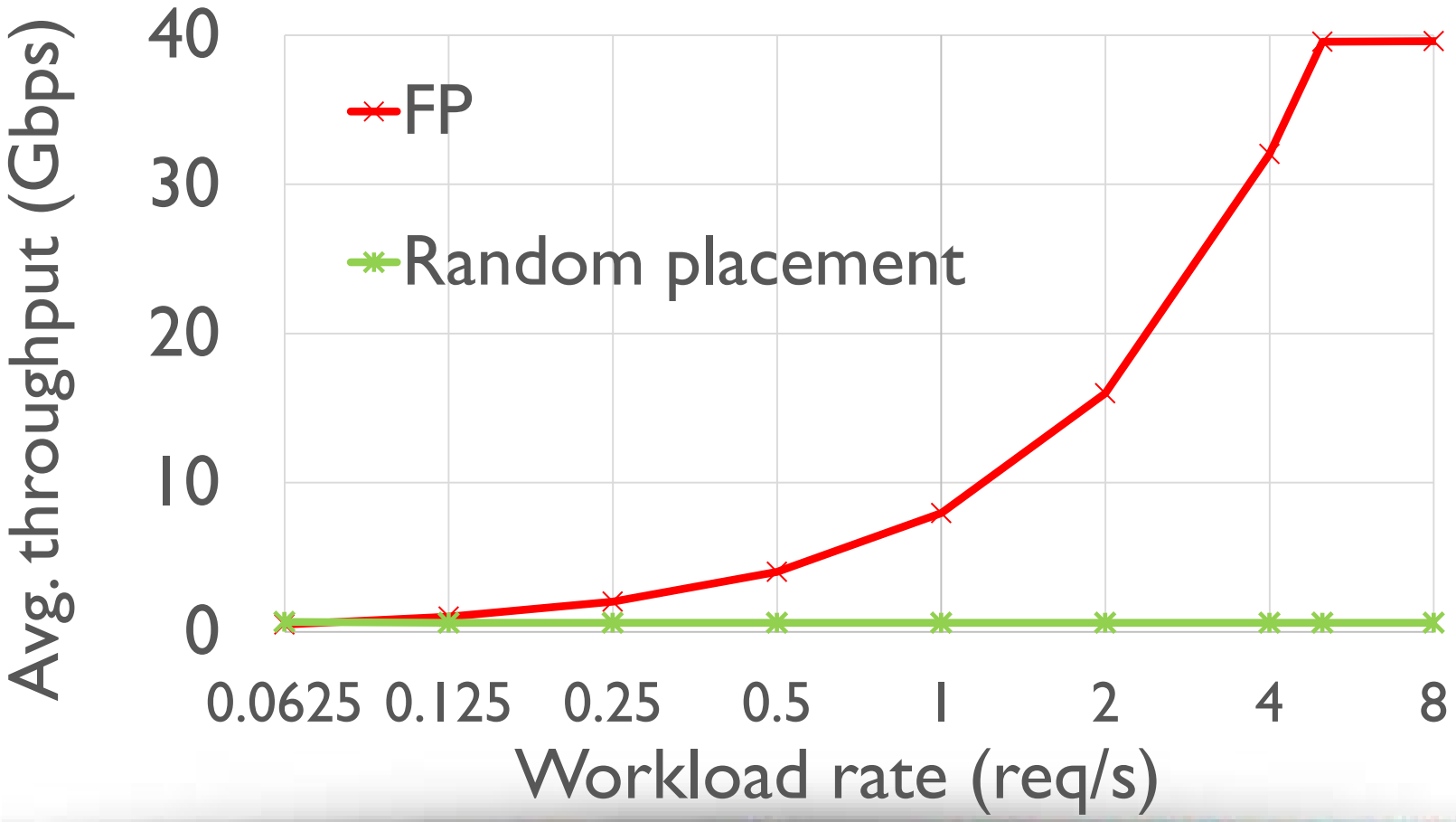
# Rack throughput






2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

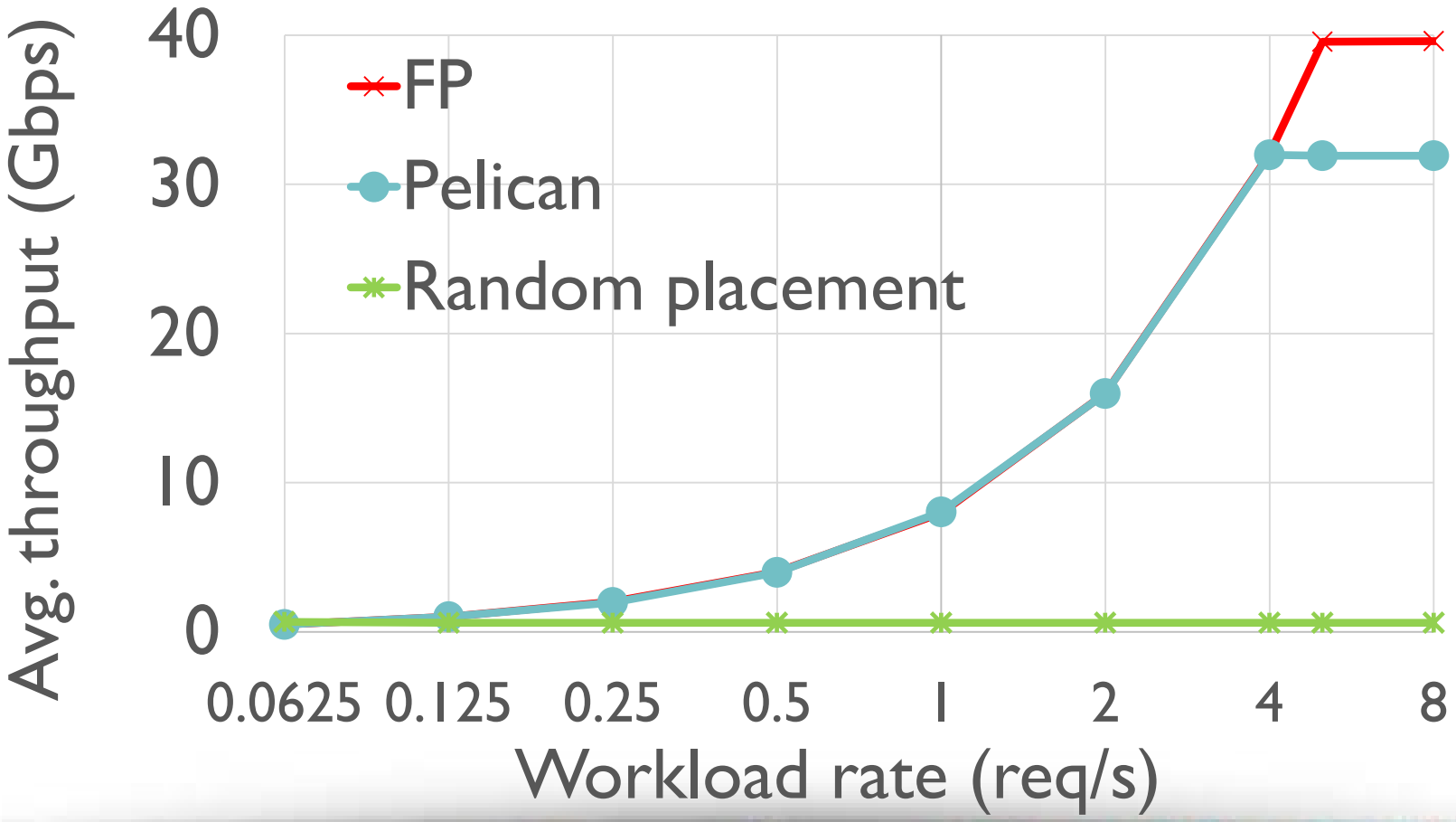
# Rack throughput






2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

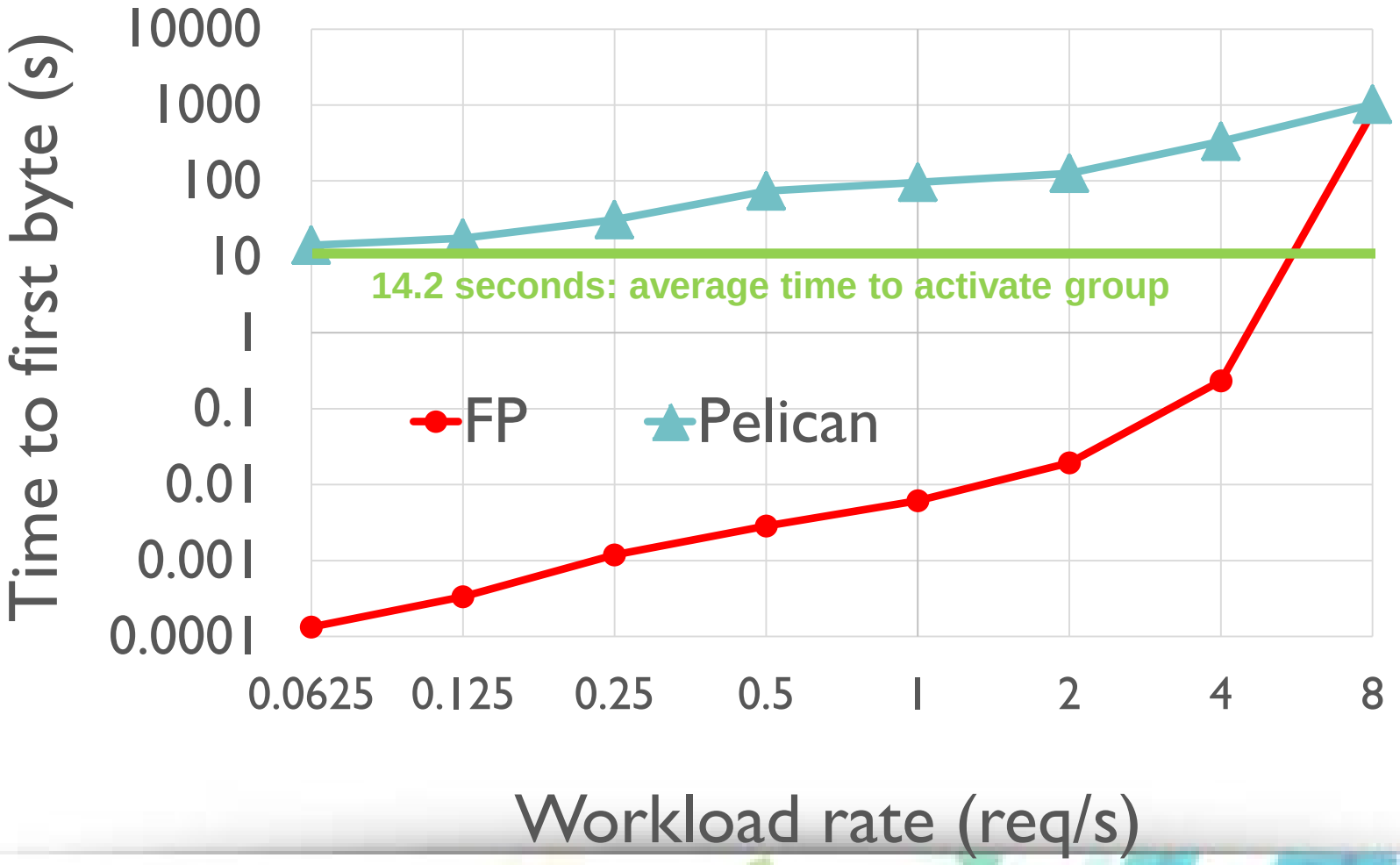
# Rack throughput






2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

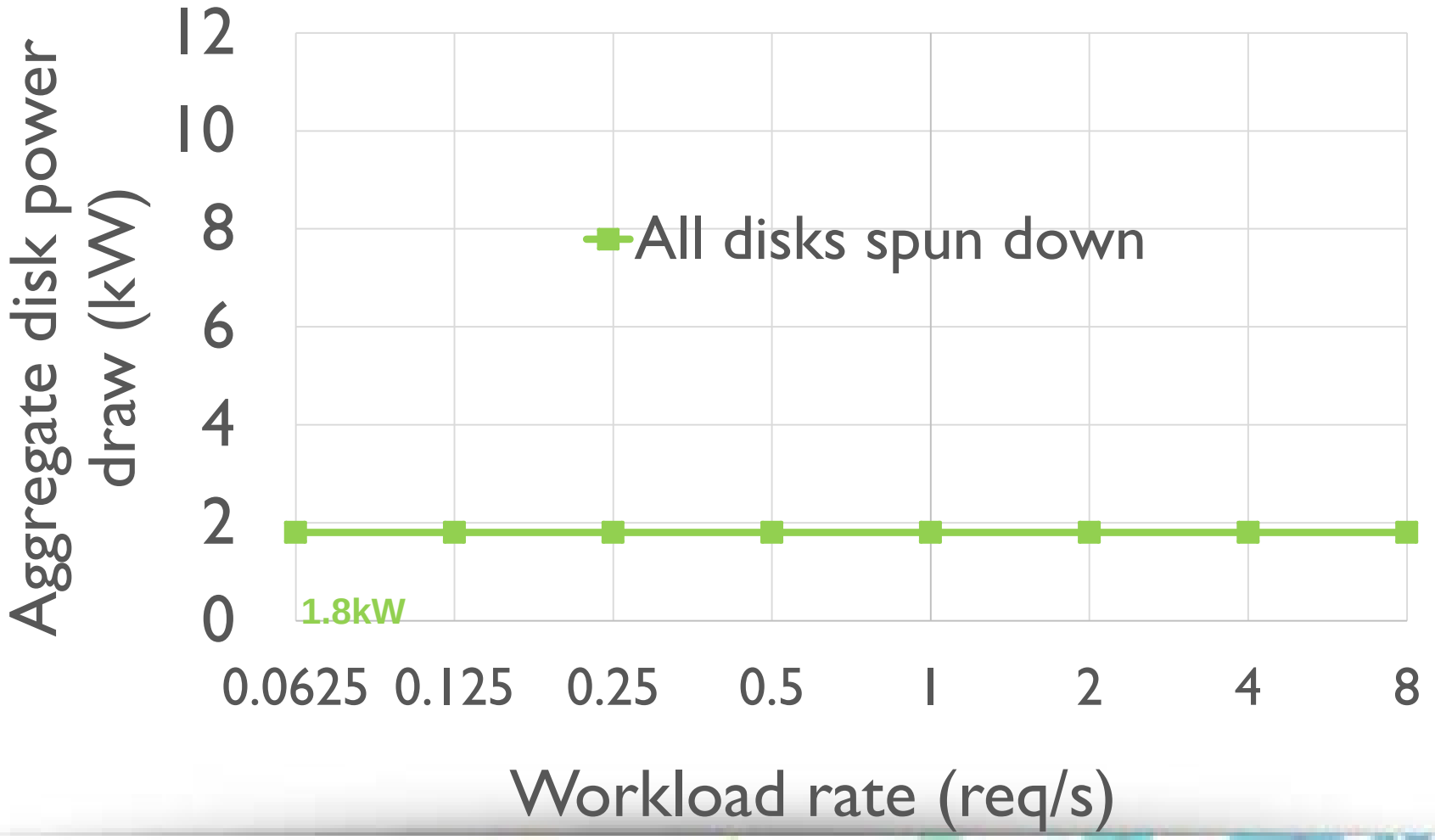
# Time to first byte






2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

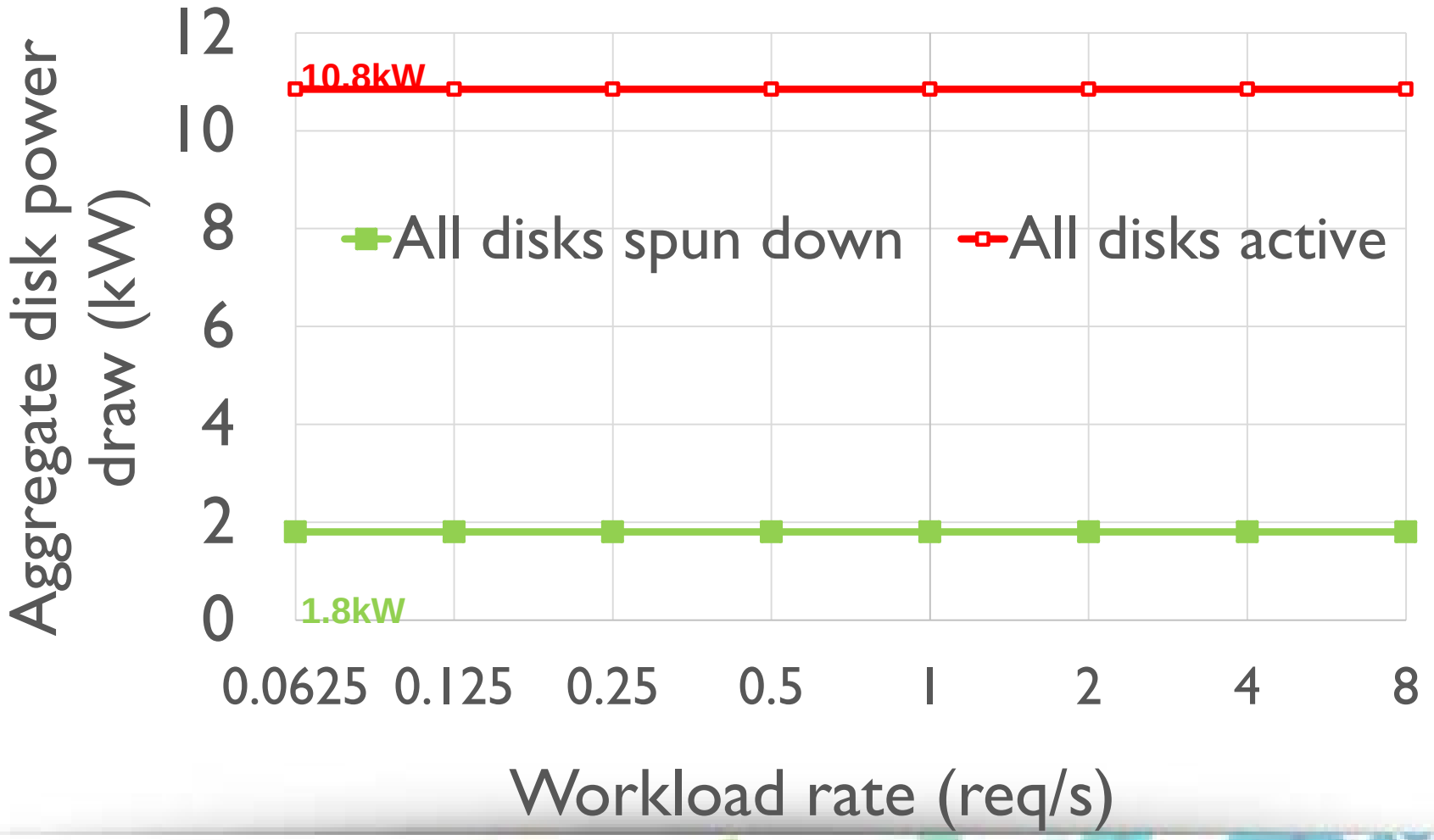
# Power consumption






2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

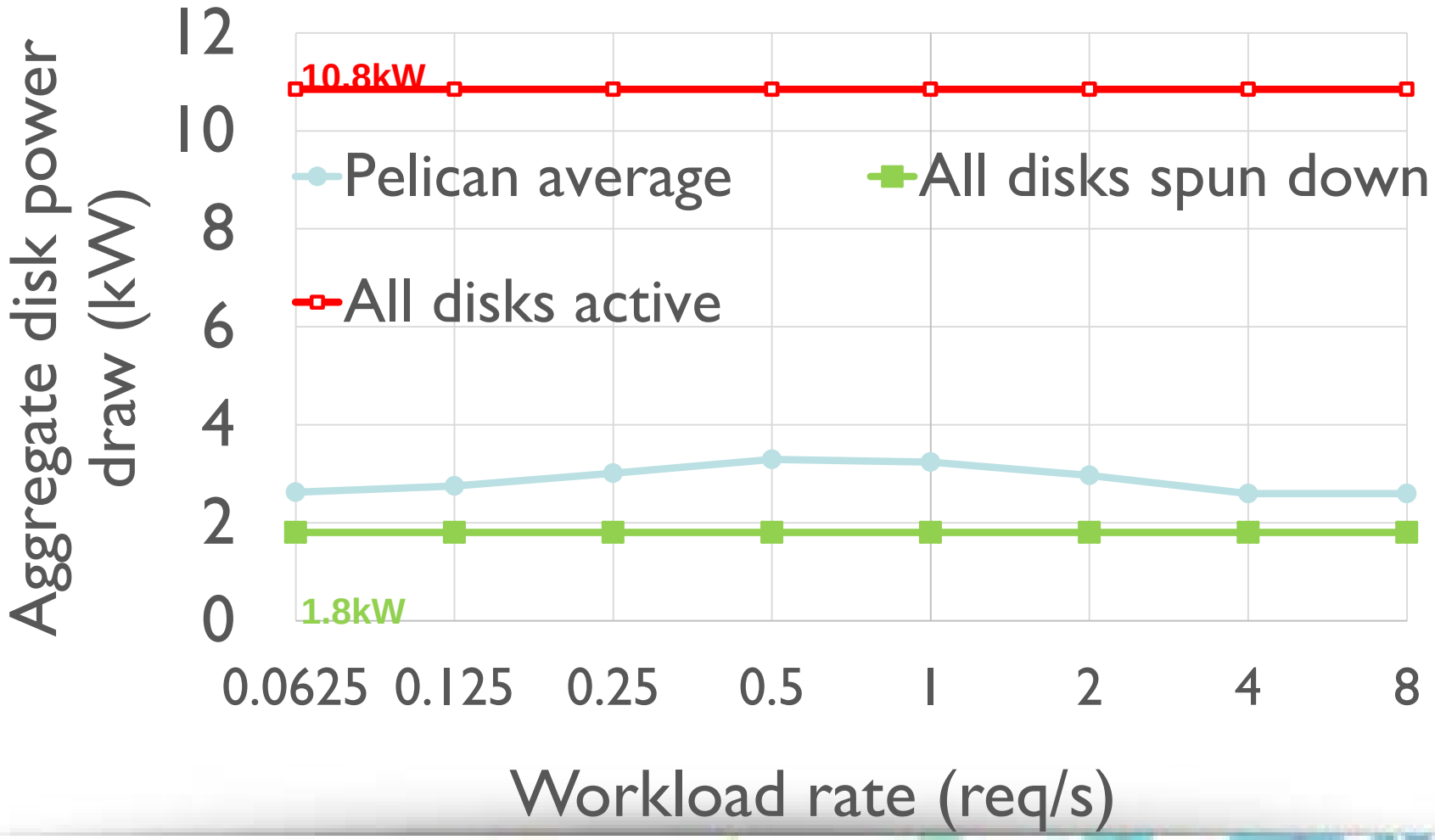
# Power consumption






2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

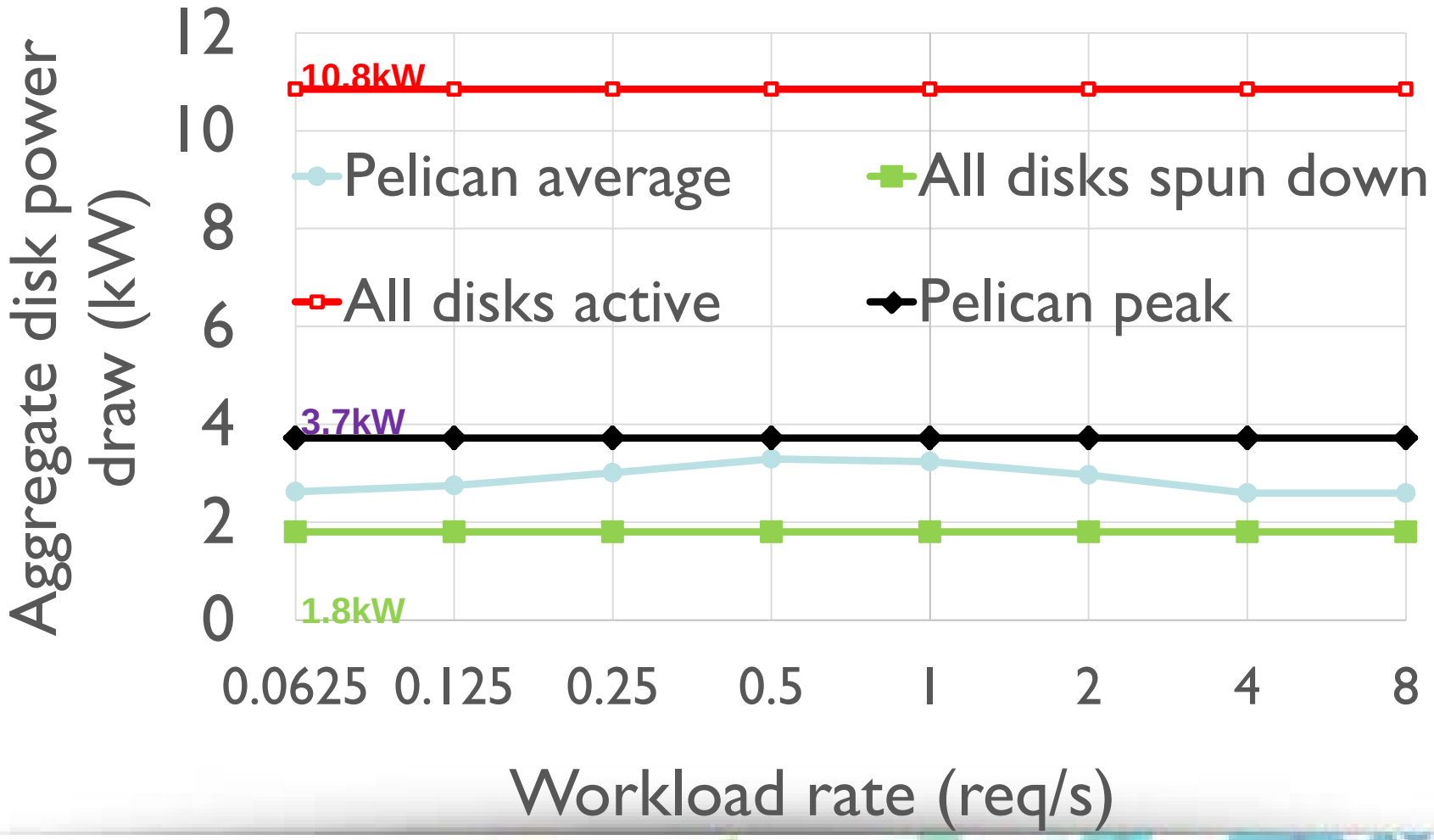
# Power consumption






2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# Power consumption: 3x lower peak





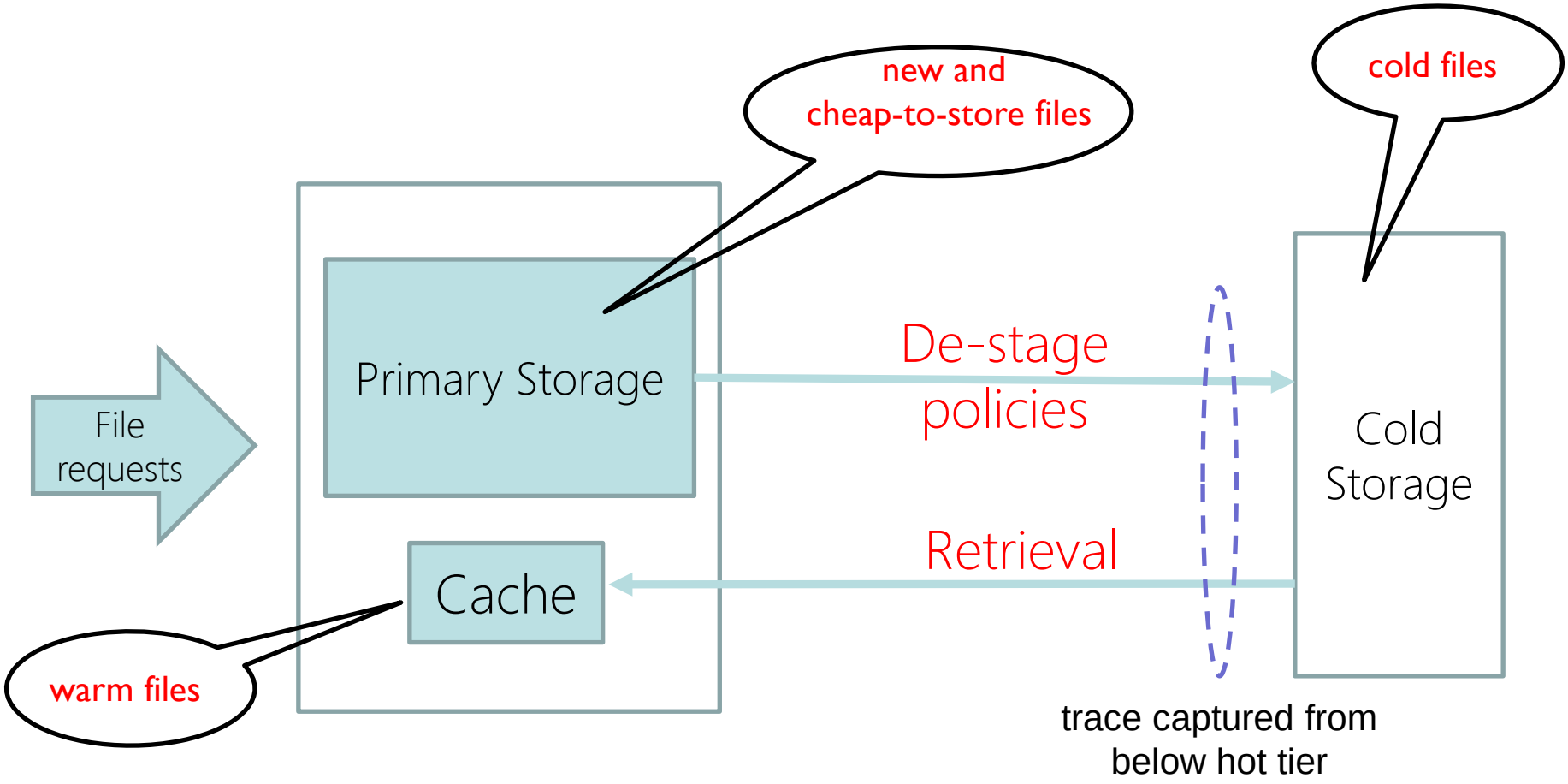

2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# Trace replay

- European Centre for Medium-range Weather Forecasts [FAST 2015]
  - ECFS trace is every request for 2.4 years.
  - Run through a tiering simulator


2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

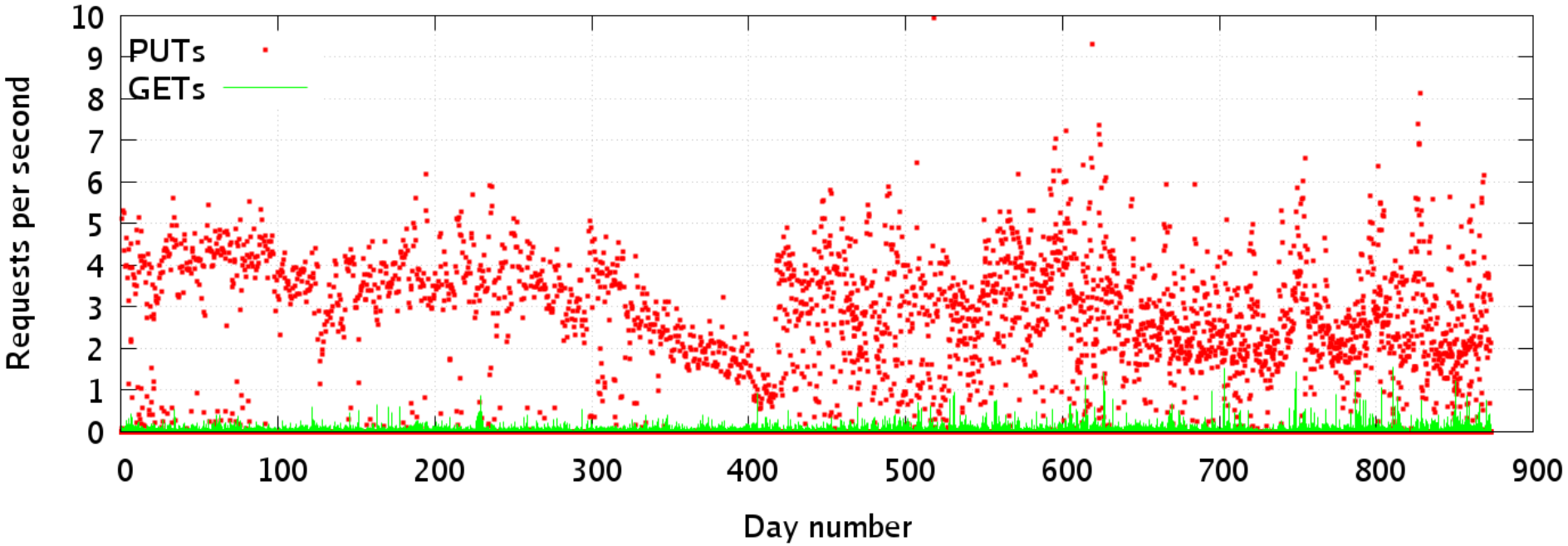
# Tiering model






2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# Requests per second, over 2.4 years

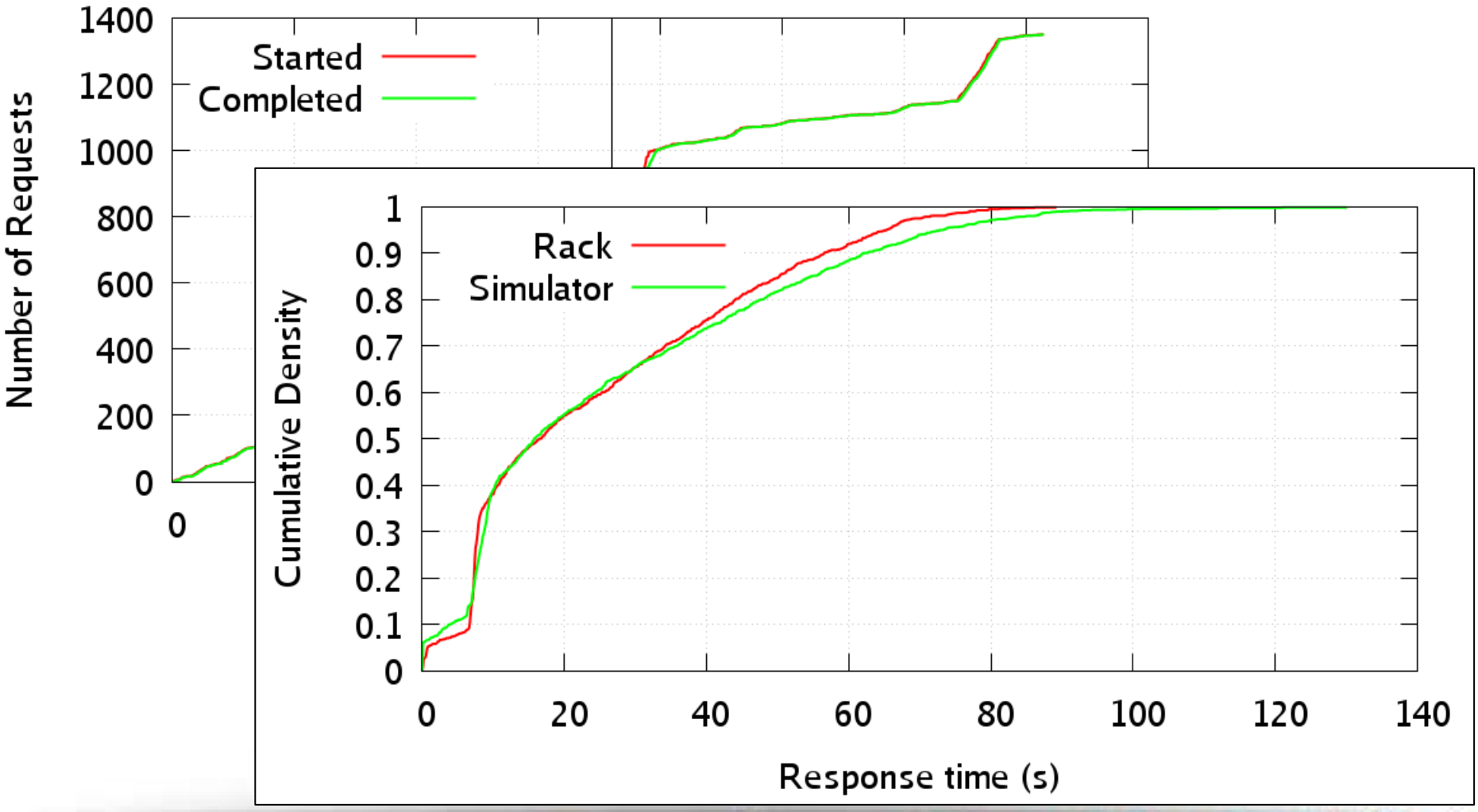



- We replay two 2-hour segments:
  - G1: highest response time
  - G2: deepest queues



2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

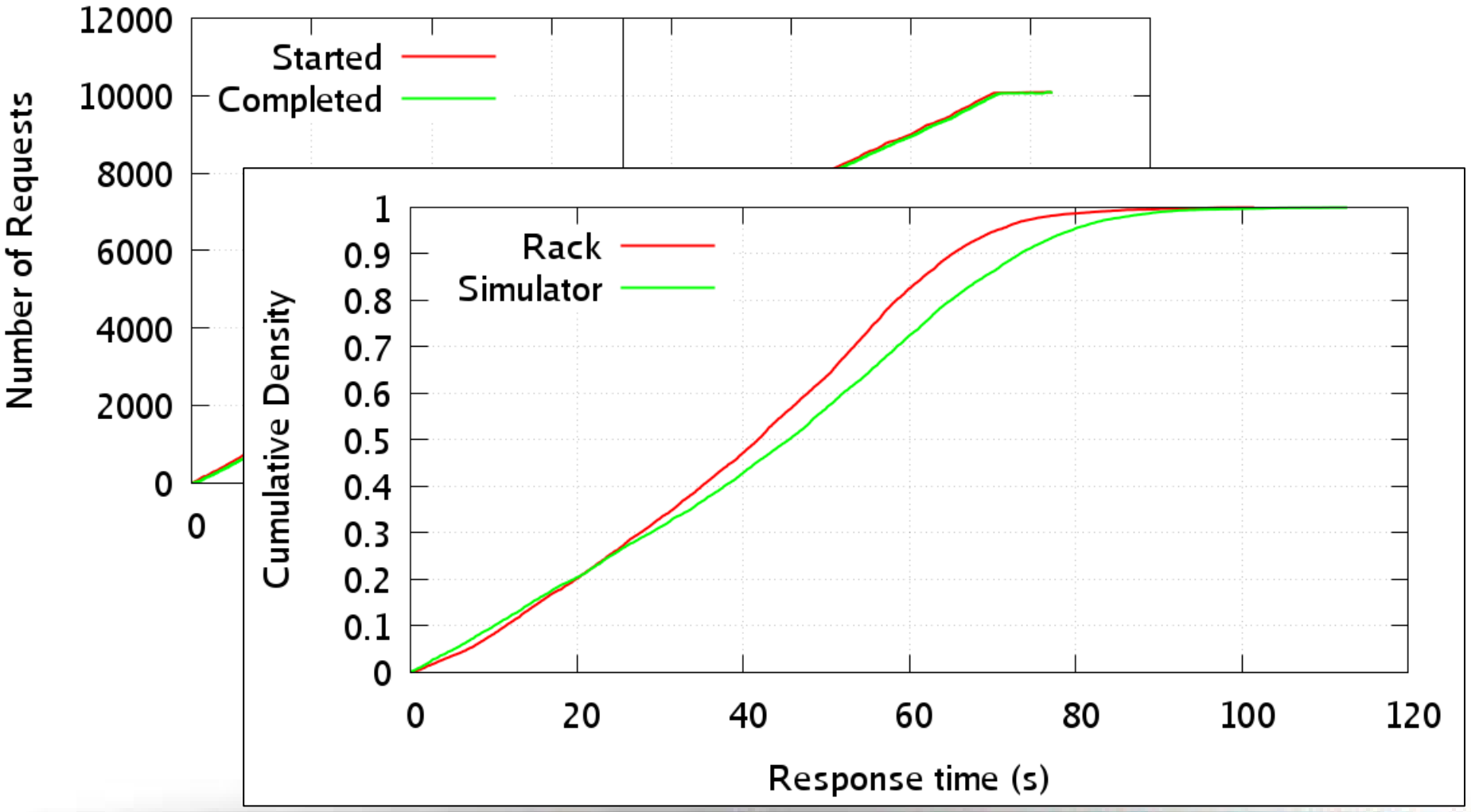
# G1: Highest response time






2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# G2: Deepest queues






2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# War stories

- Booting a system with 1152 disks
  - BIOS changes needed
- Object store vs. File system
- Data model for system:
  - Serial numbers on all FRUs
  - Disks, Volumes, Media


2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.

# Thank you!

*Questions?*




2015 Storage Developer Conference. © Microsoft Research. All Rights Reserved.