Understanding Rack-Scale Disaggregated Storage

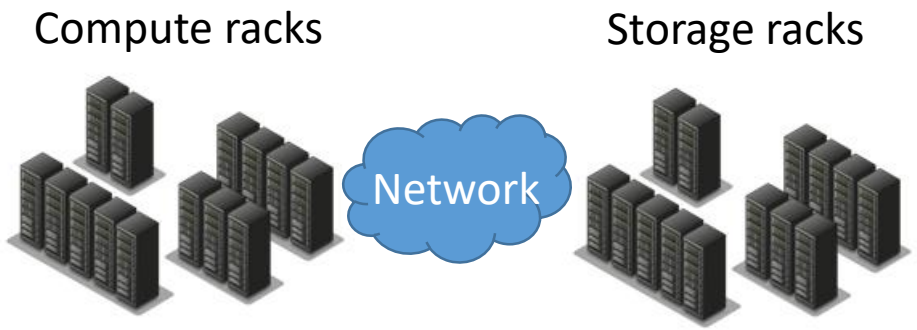
Sergey Legtchenko, Hugh Williams, Kaveh Razavi[†], Austin Donnelly, Richard Black, Andrew Douglas, Nathanael Cheriere[†], Daniel Fryer[†], Kai Mast[†], Angela Demke Brown[‡], Ana Klimovic[†], Andy Slowey and Antony Rowstron

Microsoft Research

[†]intern, [‡]visiting researcher

Storage Disaggregation

Common in the cloud:



Improves performance/cost:

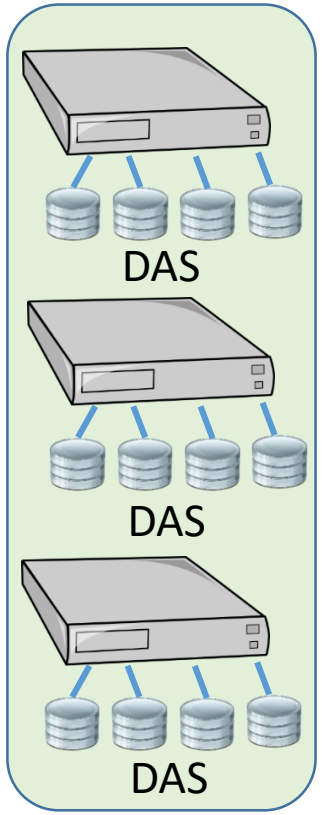
- Independent resource scaling
- Rack hardware specialization

Does not happen in HDD storage racks:

- Shared-nothing Servers
- Direct-attached Storage (DAS)

Strict HDD Ownership Principle:

- HDD **always** managed by the server to which it is physically attached



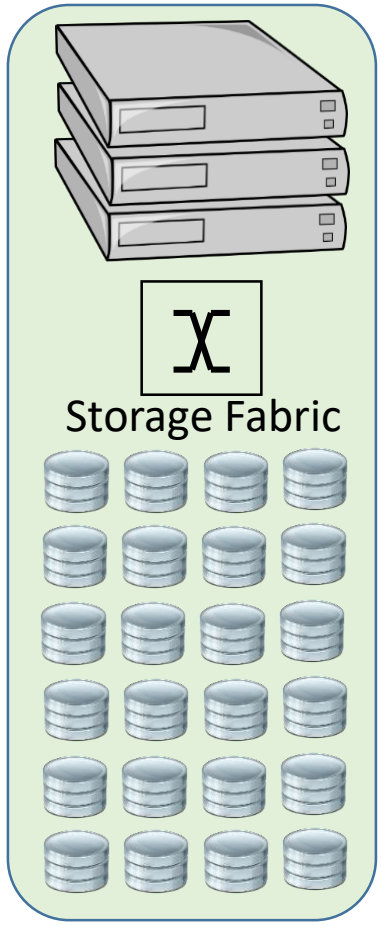
Do we need rack-scale storage disaggregation?

Rack-Scale HDD Storage Disaggregation

- Relaxing the HDD Ownership Principle
  - At a given time, a HDD is managed by one server…
  - …but it is possible to reconfigure which server it is.
- Enables 4 types of disaggregation:
  - Configuration Disaggregation
  - Failure Disaggregation
  - Dynamic Elastic Disaggregation
  - Complete Disaggregation

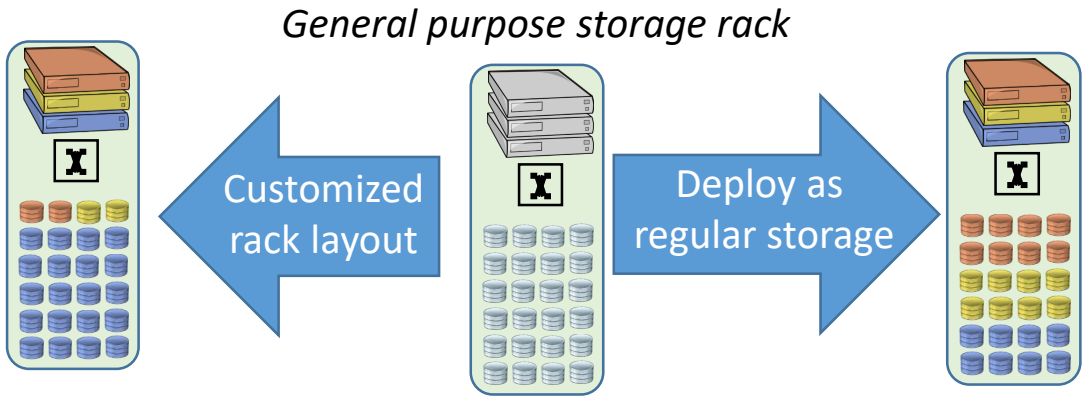
No reconfiguration during normal operation

Reconfiguration part of normal operation

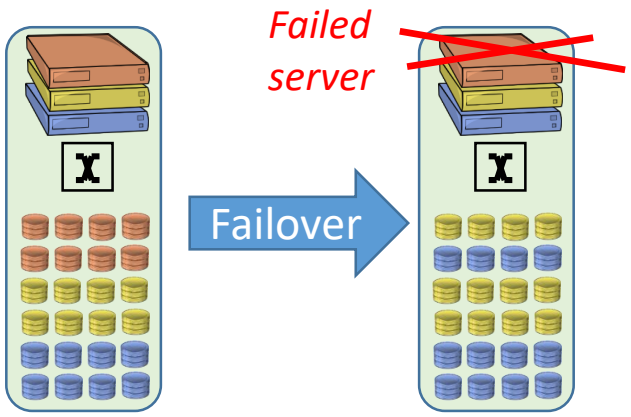


No Reconfiguration during Normal Operation

- **Configuration Disaggregation**
  - One rack for all workloads
  - Configure once at deployment
  - Optimized offline for workload

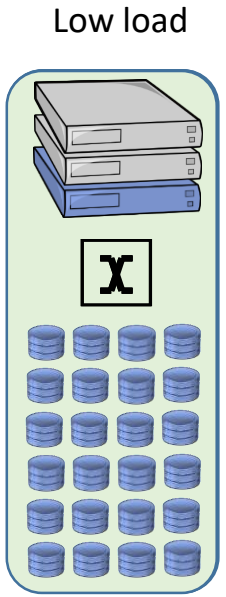


- **Failure Disaggregation**
  - Reconfigure on server failure
  - Move HDDs, not data



Reconfiguration is part of Normal Operation

- **Dynamic Elastic Disaggregation**
  - Dynamically adapt HDD-to-server ratio
  - High load: each server gets its fair share of HDDs
  - Low load: most HDDs attached to few servers



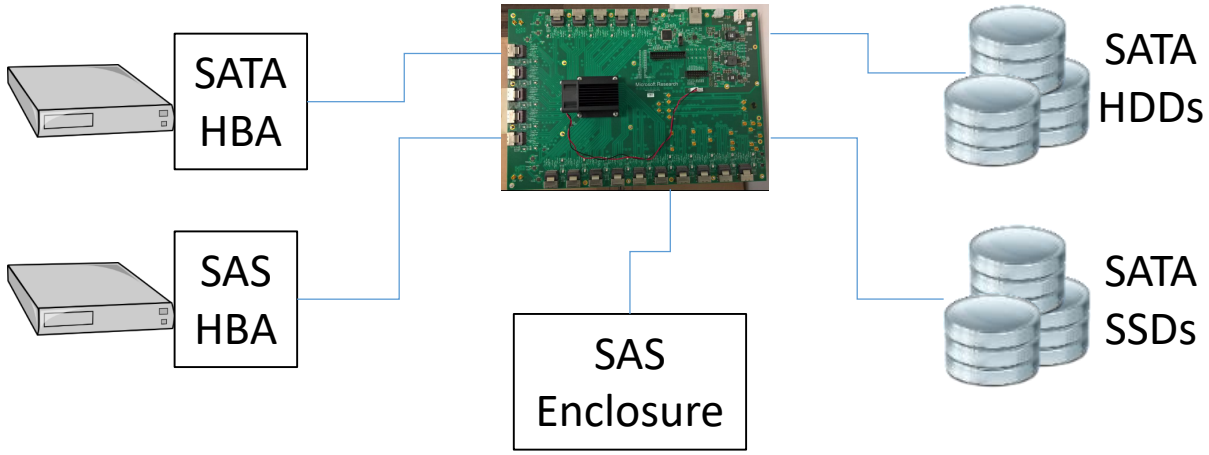
- **Complete Disaggregation**
  - Reconfigure at IO granularity
  - Any server can IO to any file on any HDD

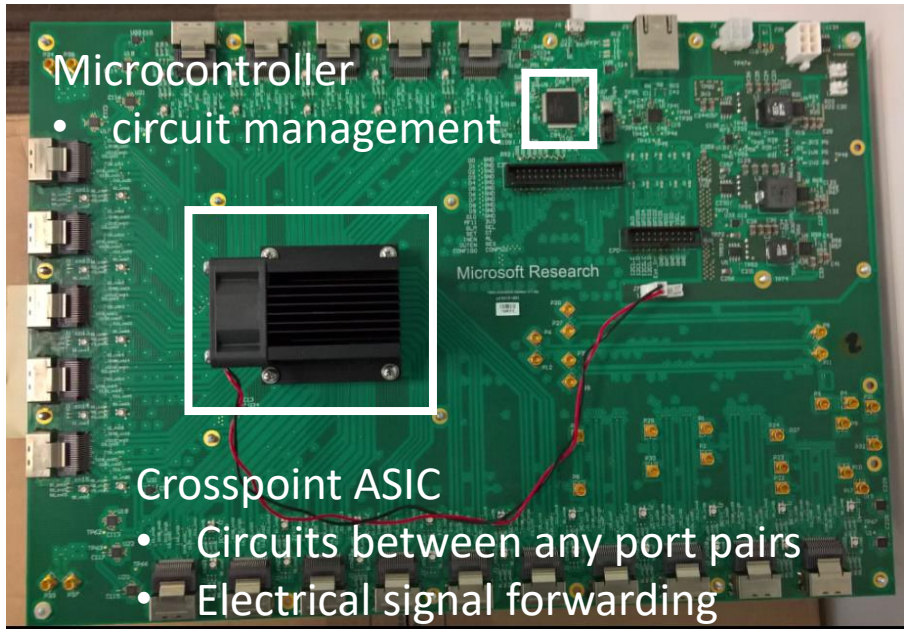
Summary of Disaggregation Scenarios

	Configuration Disaggregation	Failure disaggregation	Dynamic Elastic Disaggregation	Complete Disaggregation
Storage stack redesign	No	Small	Moderate	High
Online controller	No	Not necessarily	Yes	Yes (on IO path)
Reconfiguration frequency	O(rack lifetime)	O(server failures)	O(hours-days)	O(IO rate)
Reconfiguration overhead	None	Not under normal operation	Low	High

A Fabric to Explore Rack-Scale Disaggregation

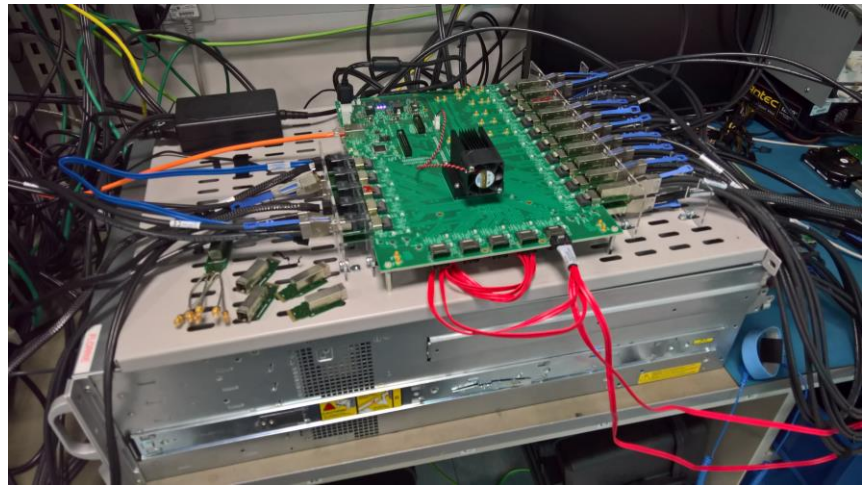
- Storage switch
  - Custom hardware design
  - Circuit switch abstraction
  - 160 ports @ 6 Gbps/port
- Benefit of the design: extreme flexibility





Experience with Configuration Disaggregation

- Easy to enable
  - No controller
  - No reconfiguration overhead
  - Unmodified software on servers
- Simplifies management & operation
  - One storage rack for all workloads
- Also very useful for development
  - **We use it on a daily basis!**
  - Fast instantiation of test configurations



Our test setup for configuration disaggregation

Experience with Failure Disaggregation

- Hardware trends impact data availability:
  - HDD and SSD capacities grow
  - Servers can have a LOT of direct-attached storage
  - e.g.: **Petabytes** of data per Pelican (cold storage) server
  - On failure, amount of data and time to recover increases
- Failure disaggregation improves availability
  - Reduces data unavailability to tens of seconds or less
  - No resources used to rebuild data
  - No reconfiguration overhead for normal operation

Pelican prototype has:

- **1152 HDDs/rack**
- **2 servers**

Exploring Dynamic Elastic Disaggregation

- Ongoing work
- Storage workloads are bursty
  - Average server utilization is low
  - Load skew across servers
- Online controller
  - Monitors storage traffic in the rack
  - Adapts HDD-to-server ratio
  - Not on the data path
- Better server utilization
  - Allows storage tiering within the rack
  - Some servers can host background jobs, spot VM instances

Dynamic elastic disaggregation setup

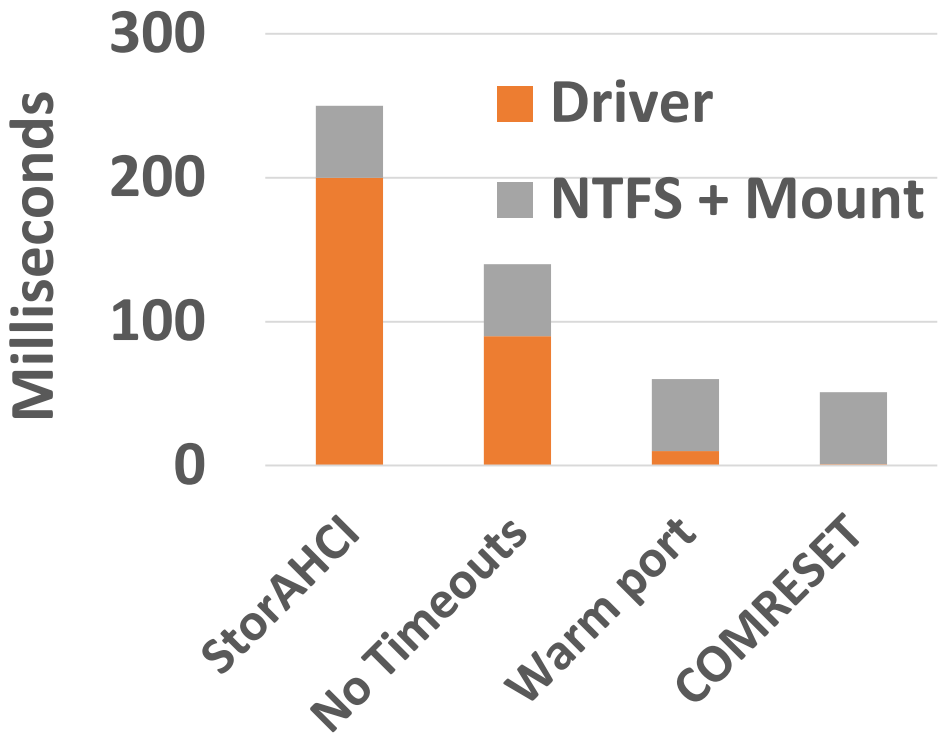
Storage fabric

SAS enclosures

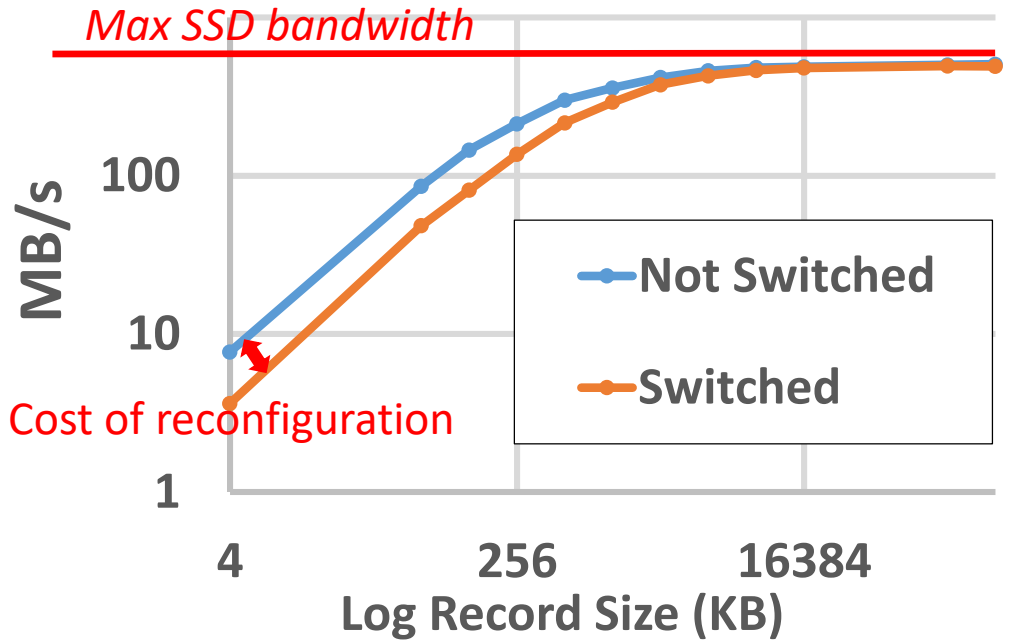
Complete Disaggregation, seriously?

- Can we reconfigure per IO?

Time to switch and mount SATA SSD:

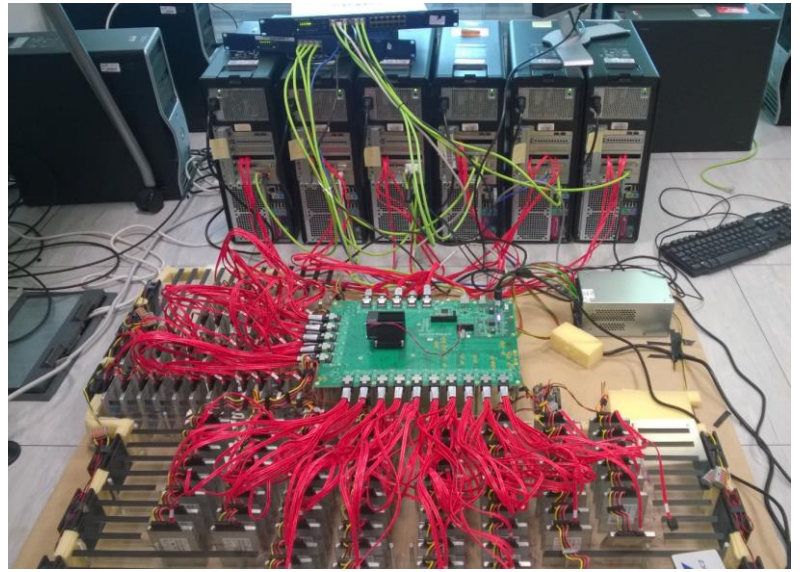


Impact on throughput of switching after every IO: (no File system mount)



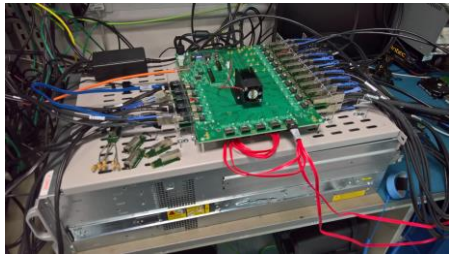
Experience with Complete Disaggregation

- A lot of pain:
  - Ecosystem challenges
  - Redesign of the storage stack
  - High overhead for small IO
  - Meta data service on the IO path
  - Hard to implement/debug
- Benefits
  - Fine-grain load balancing
  - Server failure tolerance by design



Complete disaggregation setup
- 120 SATA SSDs
- 4 servers, 3 SATA ports/server

Conclusion



- In the cloud today: no disaggregation in storage racks
  - Fixed drive-to-server mapping
- We designed a storage fabric to explore in-rack disaggregation
- Rack-scale storage disaggregation can be useful and affordable
  - Configuration disaggregation
  - Failure disaggregation
  - Dynamic elastic disaggregation

  → Substantial benefits; No/small reconfiguration overheads; Little or no software/hardware changes

- Can become a challenge
  - Complete disaggregation → High reconfiguration overhead; Hard to implement and maintain