Optimizing Storage and Memory Hierarchies

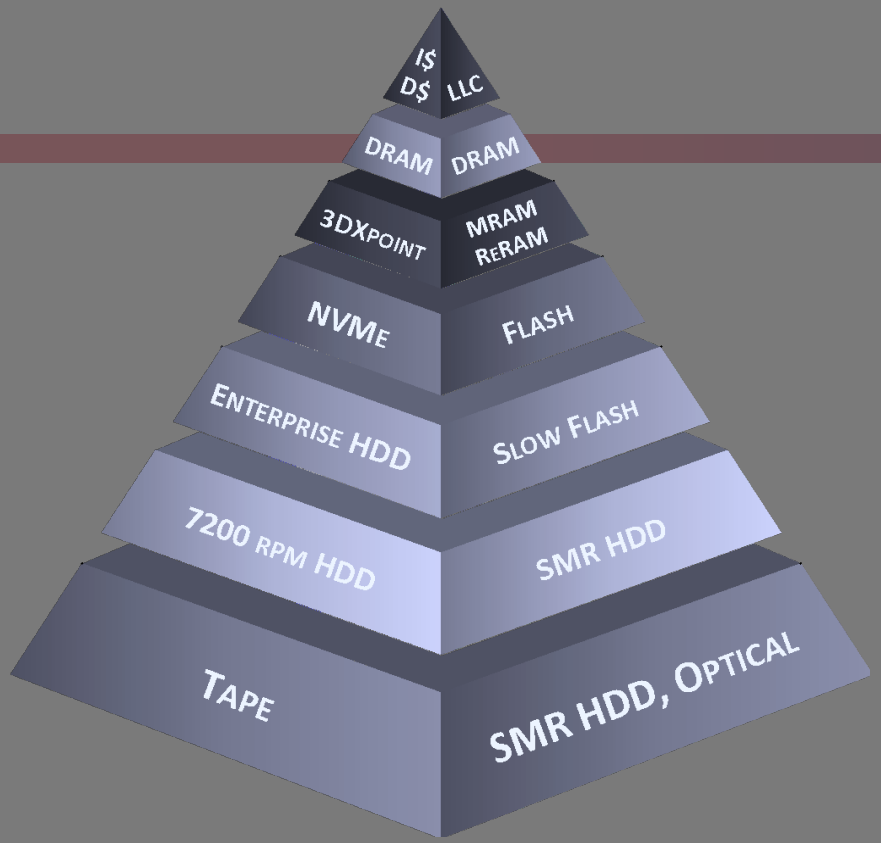
REGIONAL SDC 24

BY Developers FOR Developers

APRIL 24, AUSTIN, TX

Andy Banta – Storage Janitor
Powered by Magnition

A SNIA Event



The Challenge

Modern compute and storage system use multiple layers interacting in multiple ways

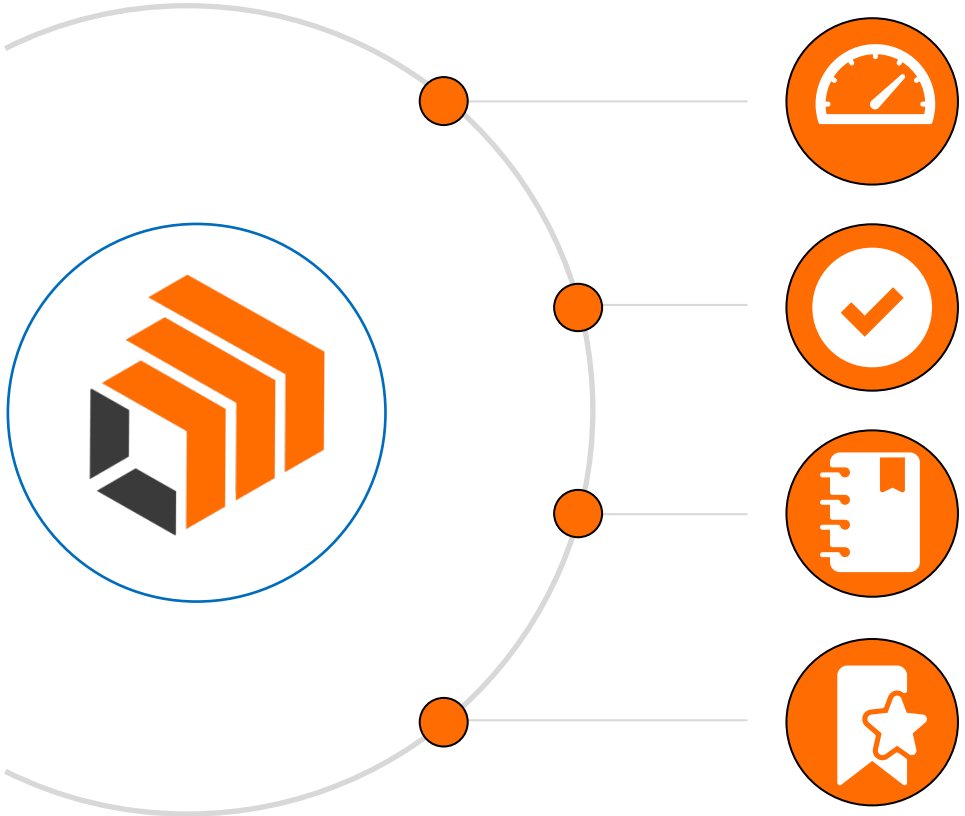
HOW CAN CURRENT TECHNOLOGY ACHIEVE…

- Latency control
- Multi-tenant thrash remediation
- Correct tier sizing
- Workload-awareness
- Hot working set management
- Latency and throughput SLAs
- Memory capacity planning

AS MORE HARDWARE LAYERS ADD COMPLEXITY?

ABOUT MAGNITION

STORAGE PERFORMANCE, REINVENTED



World's First Real-Time Data Placement Optimization

Patented technology is a first for the industry.

Proven At-Scale, with Production Workloads

Use customer traces to fully test diverse workloads in real-time.

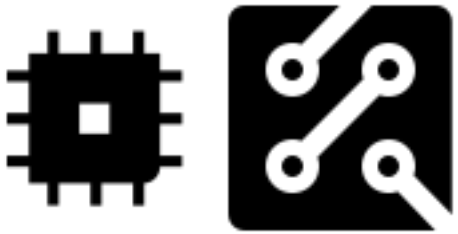
Peer-Reviewed and Published in Leading Journals

Multiple industry articles published and reviewed.

Award-Winning, Patented Technology

3-time award winner for innovative technology.

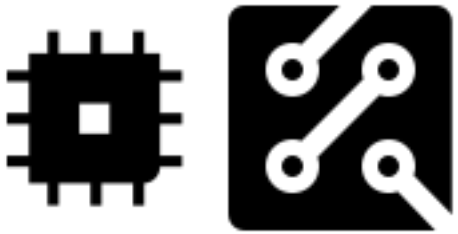
Engineering simulation



REGIONAL SDC 24

Engineering simulation

- Cheaper, faster, more flexible than system building
- Engineering design uses simulations, why not software?



REGIONAL SDC 24

Value of simulations

Faster and easier to prototype

Minimal up-front hardware costs

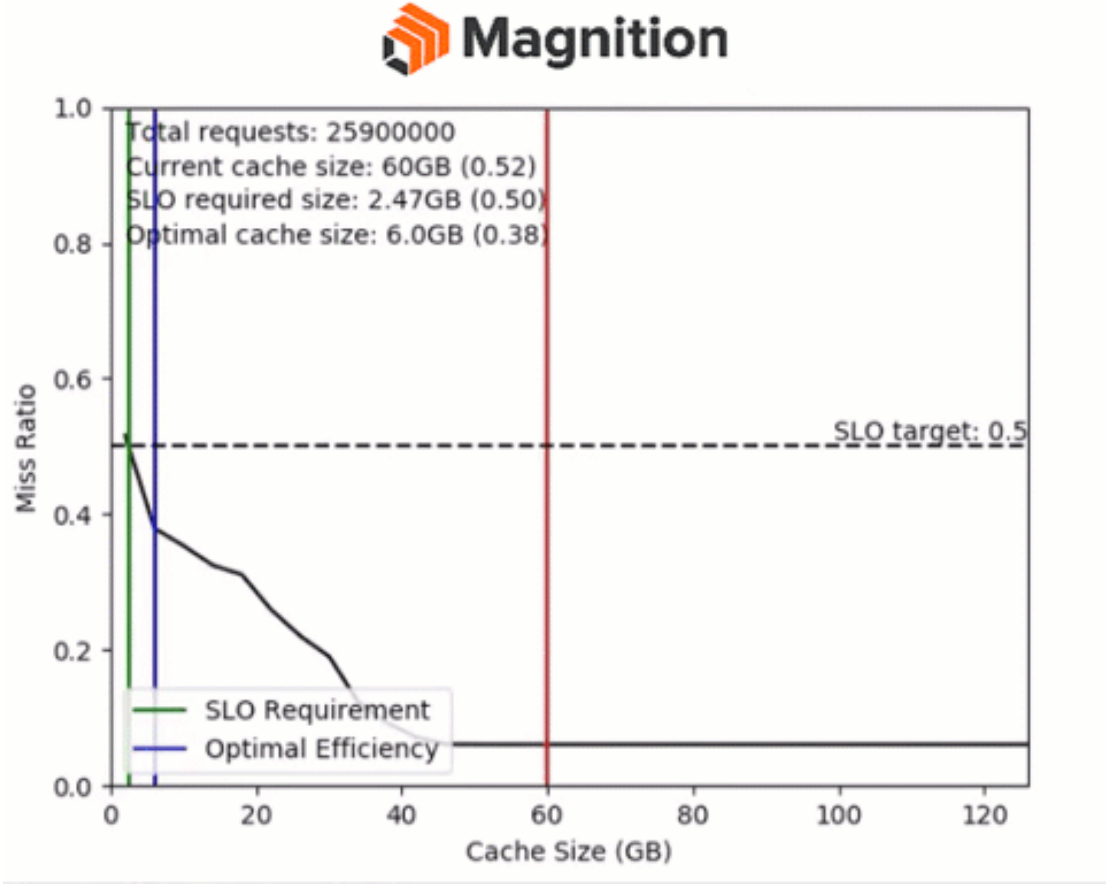
Great opportunities for optimizations

Loads of simulations are done at ASIC level

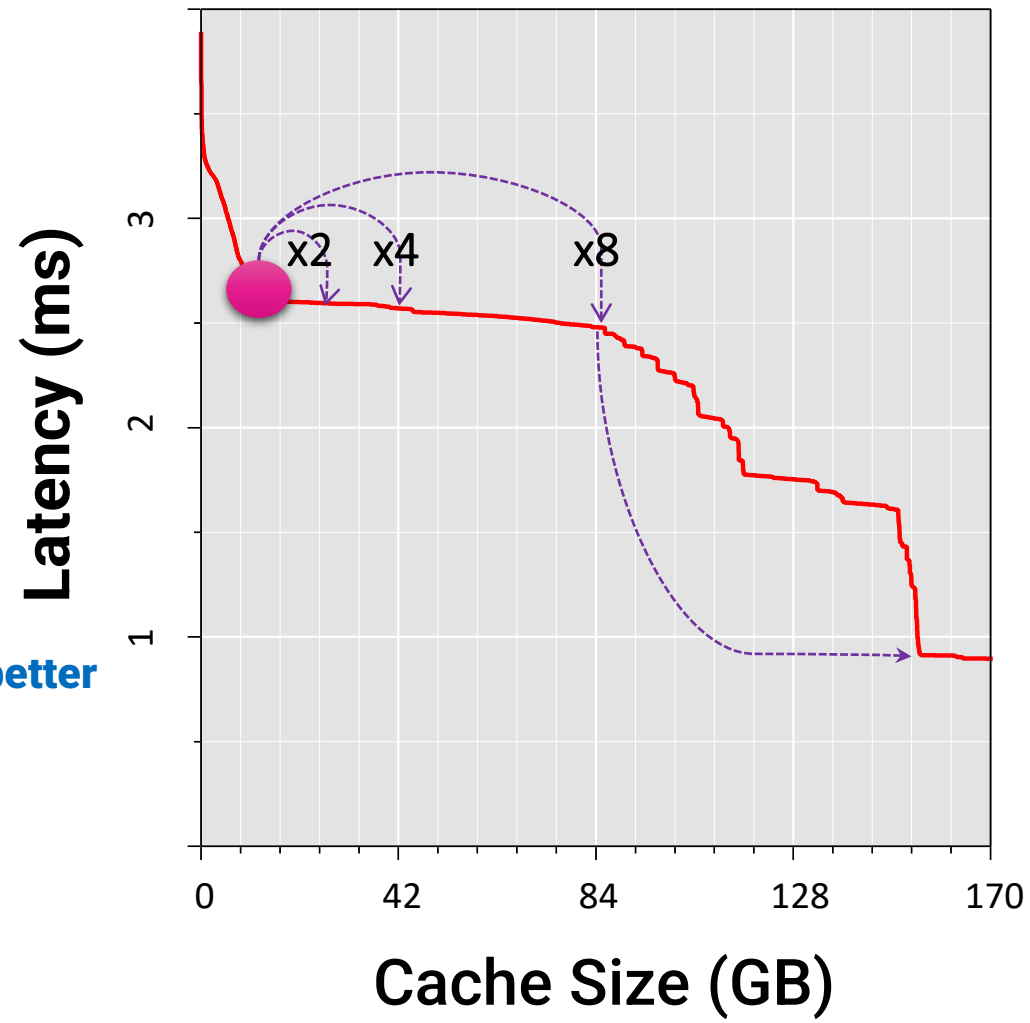
> The same practices should apply to component and software levels

Choose three

1. Lower cost
2. Higher speed
3. More flexibility



PERFORMANCE VERSUS RESOURCE ALLOCATION



Models help decide useful increments of change

In this example, no benefit despite an 8x increase in budget

A different approach to optimization

Compose simulations of complex memory and storage

Break the simulation into components

Allows the components to be assembled like building blocks

Provide reasonable but constrained set of variables

Run simulations with synthetic data or actual IO traces

Composable components

- Provide a framework to connect components
  - Lingua Franca provides this
  - Reactors represent system pieces
- Library of components ready to use
- Allows clients to build their own modules
- Basic set of building blocks
  - Cache
  - Media
  - Wire

Composable components

- Provide a framework to connect components
  - Lingua Franca provides this
  - Reactors represent system pieces
- Library of components ready to use
- Allows clients to build their own modules
- Basic set of building blocks
  - Cache
  - Media
  - Wire

Composable components

Provide a framework to connect components

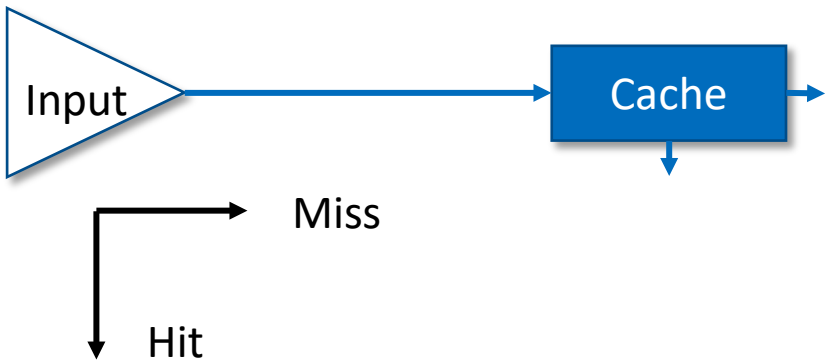
- Lingua Franca provides this
- Reactors represent system pieces

Library of components ready to use

Allows clients to build their own modules

Basic set of building blocks

- Cache
- Media
- Wire



Composable components

Provide a framework to connect components

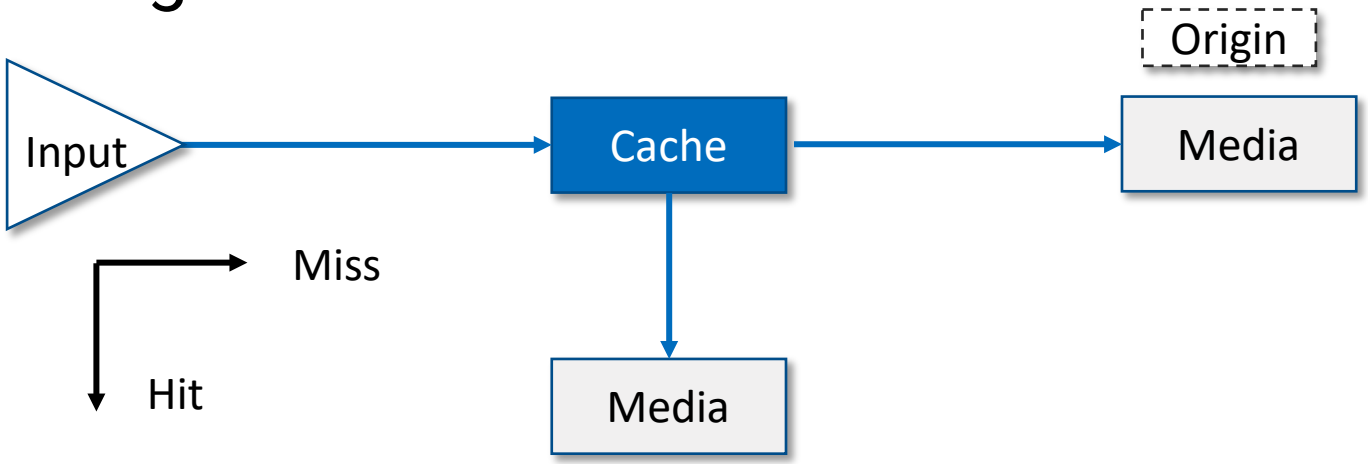
- Lingua Franca provides this
- Reactors represent system pieces

Library of components ready to use

Allows clients to build their own modules

Basic set of building blocks

- Cache
- Media
- Wire



Composable components

Provide a framework to connect components

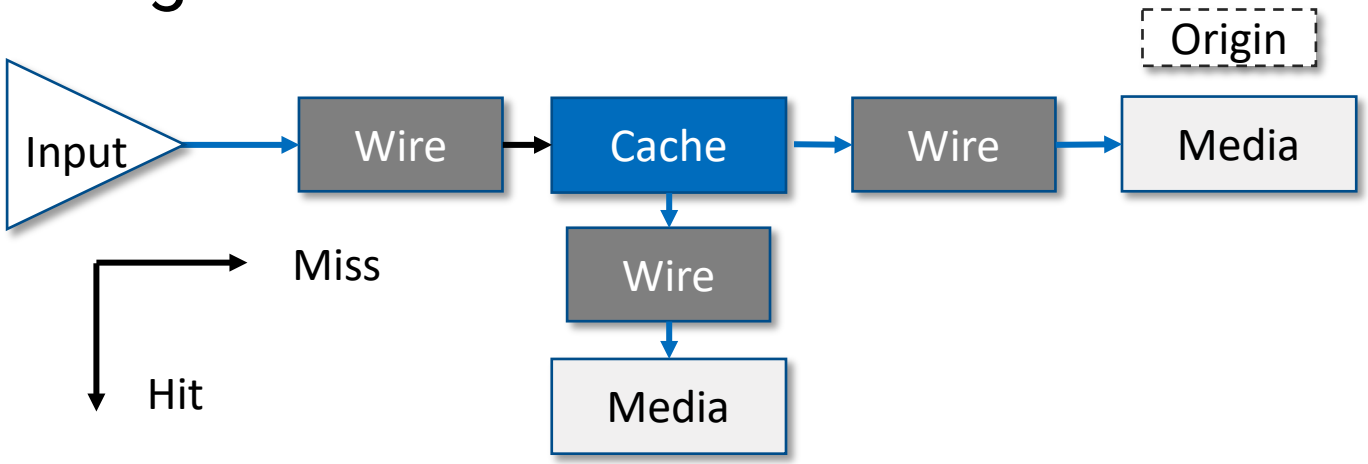
- Lingua Franca provides this
- Reactors represent system pieces

Library of components ready to use

Allows clients to build their own modules

Basic set of building blocks

- Cache
- Media
- Wire



Composable components

Provide a framework to connect components

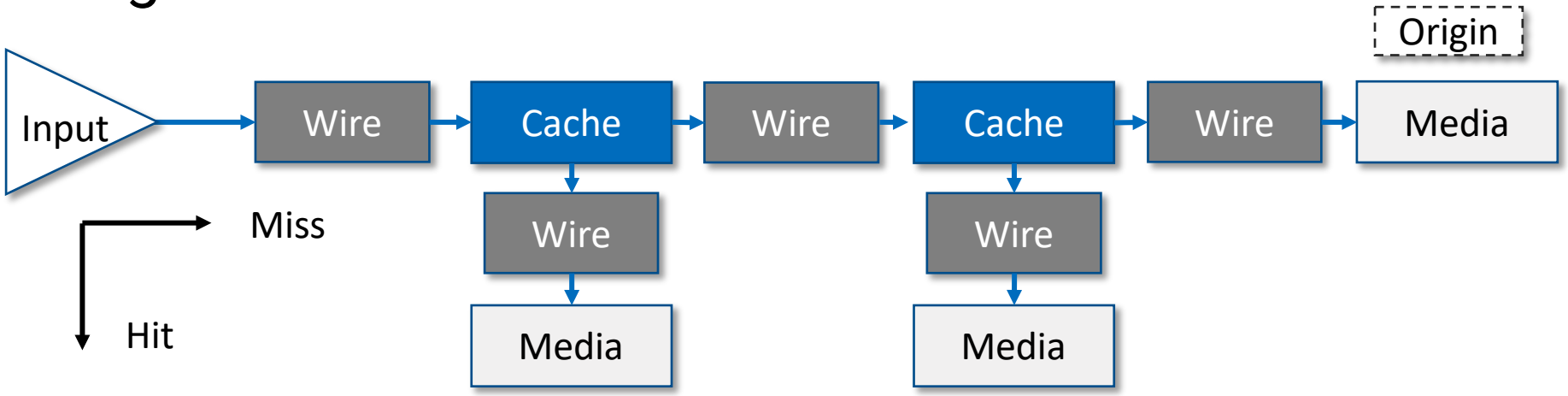
- Lingua Franca provides this
- Reactors represent system pieces

Library of components ready to use

Allows clients to build their own modules

Basic set of building blocks

- Cache
- Media
- Wire



Media component

Memory, disk, cloud storage

Introduce distinct delays

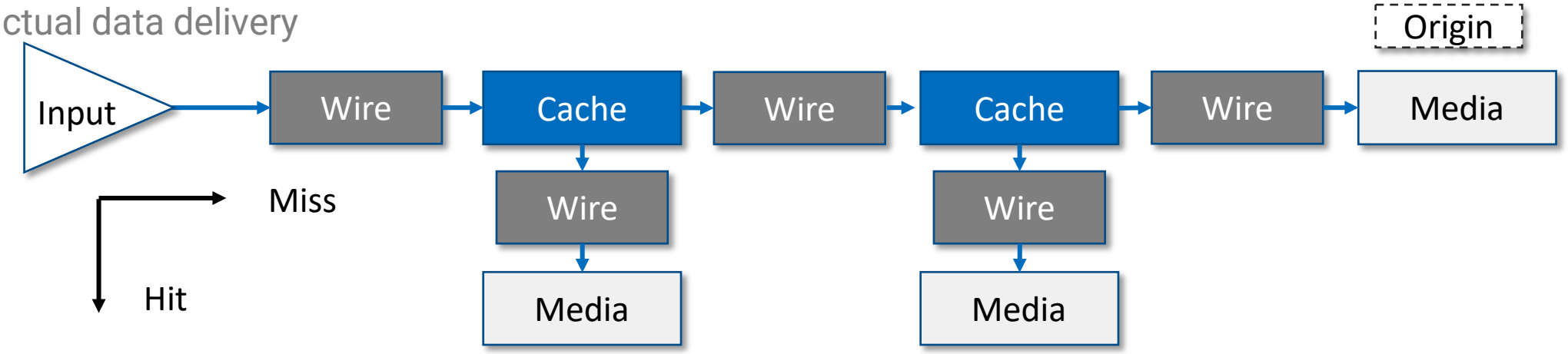
- MQSim

Parallel access

- Contention delays
- Queueing

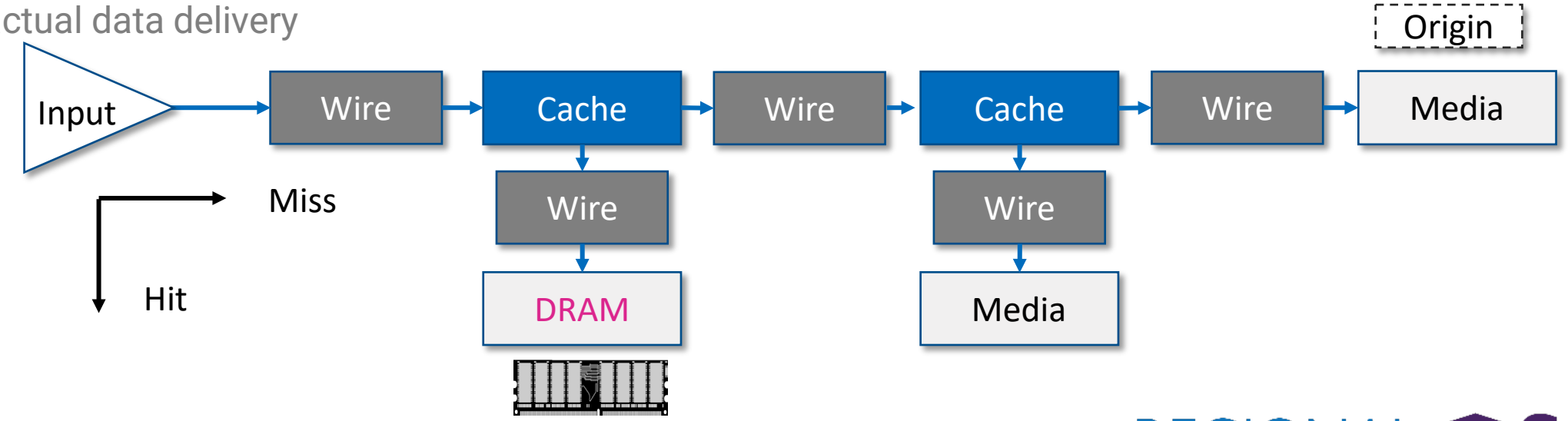
Only need to simulate delay

- Not actual data delivery



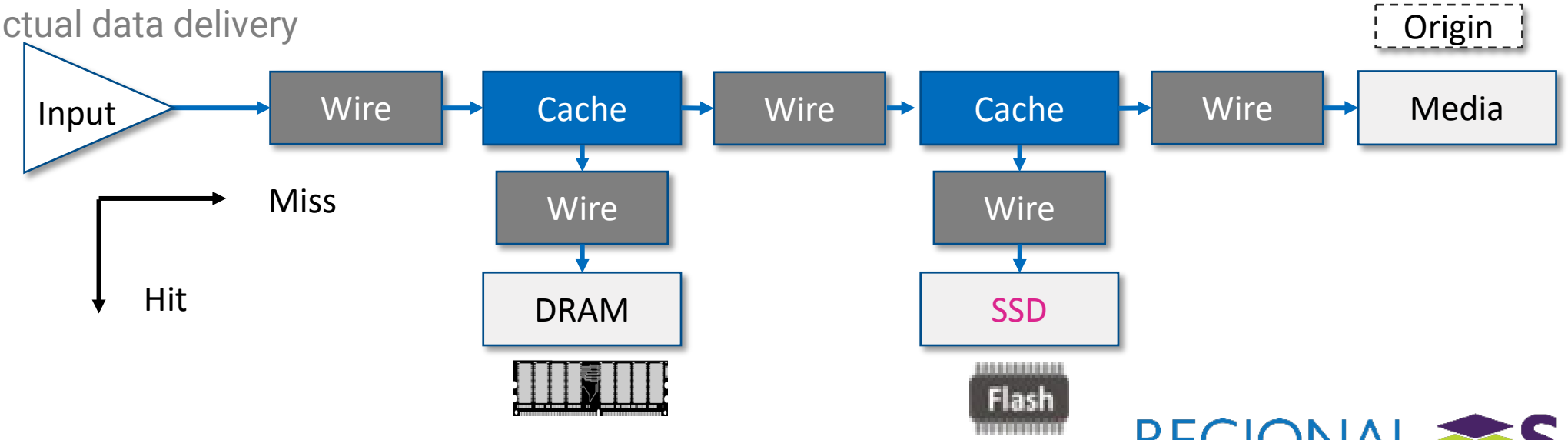
Media component

- Memory, disk, cloud storage
- Introduce distinct delays
  - MQSim
- Parallel access
  - Contention delays
  - Queueing
- Only need to simulate delay
  - Not actual data delivery



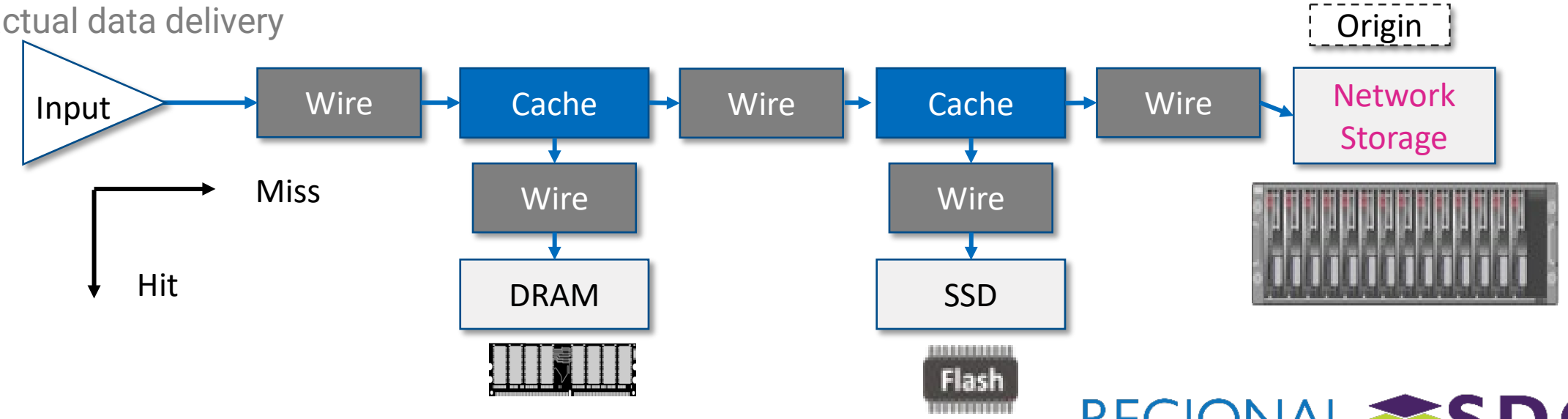
Media component

- Memory, disk, cloud storage
- Introduce distinct delays
  - MQSim
- Parallel access
  - Contention delays
  - Queueing
- Only need to simulate delay
  - Not actual data delivery



Media component

- Memory, disk, cloud storage
- Introduce distinct delays
  - MQSim
- Parallel access
  - Contention delays
  - Queueing
- Only need to simulate delay
  - Not actual data delivery



Wire component

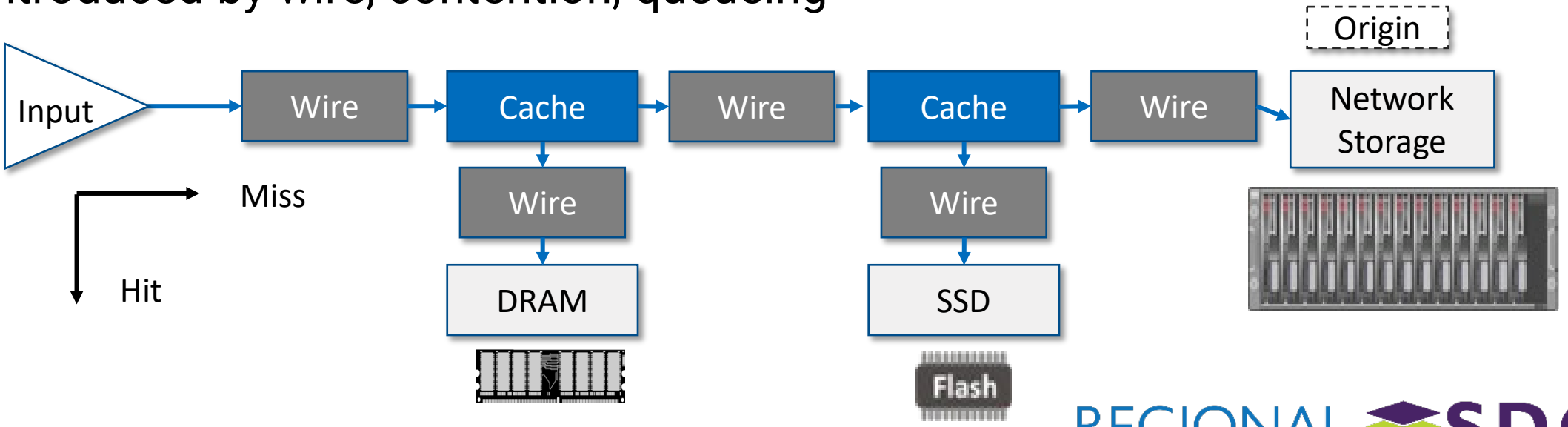
Memory bus, disk controller, network

Can multiplex and change form of IO request

Even type of wire can be variable

- Type of memory bus
- Hops in network topology

Delays introduced by wire, contention, queueing



Wire component

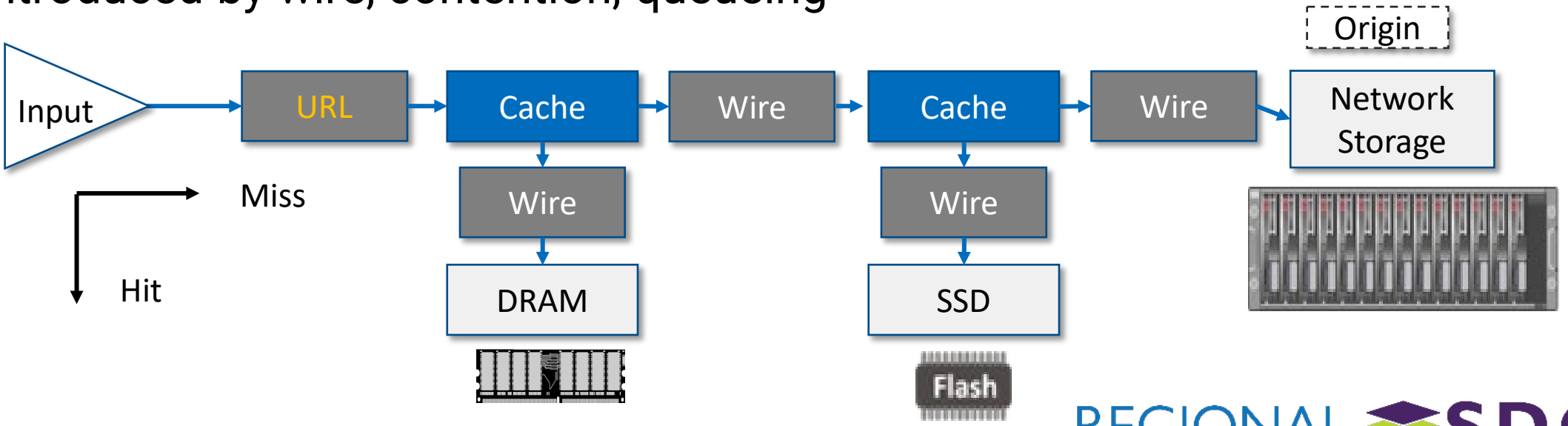
Memory bus, disk controller, network

Can multiplex and change form of IO request

Even type of wire can be variable

- Type of memory bus
- Hops in network topology

Delays introduced by wire, contention, queueing



Wire component

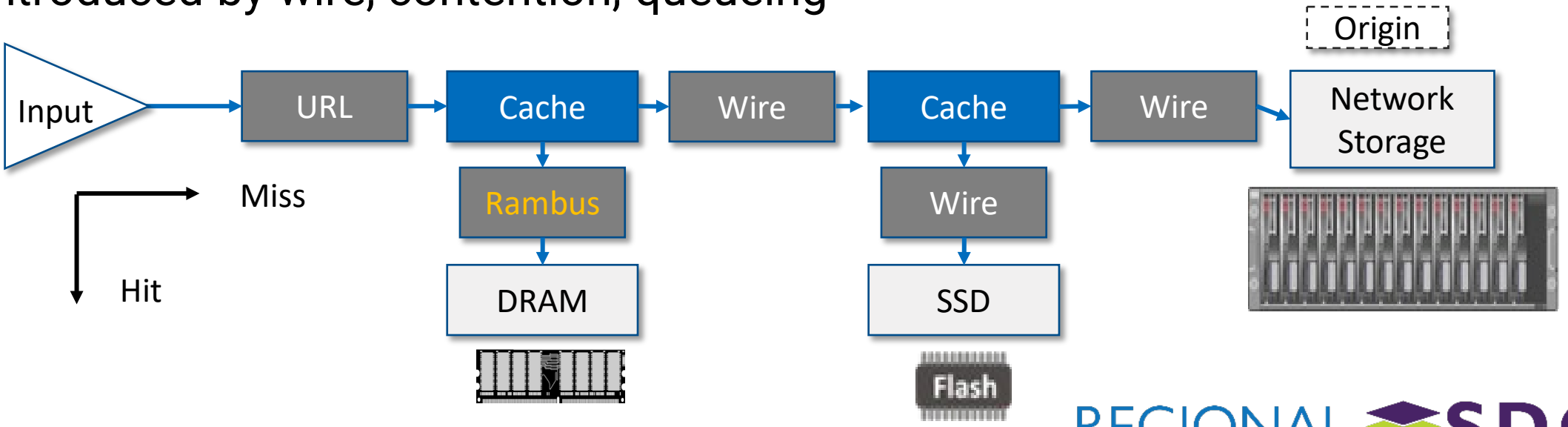
Memory bus, disk controller, network

Can multiplex and change form of IO request

Even type of wire can be variable

- Type of memory bus
- Hops in network topology

Delays introduced by wire, contention, queueing



Wire component

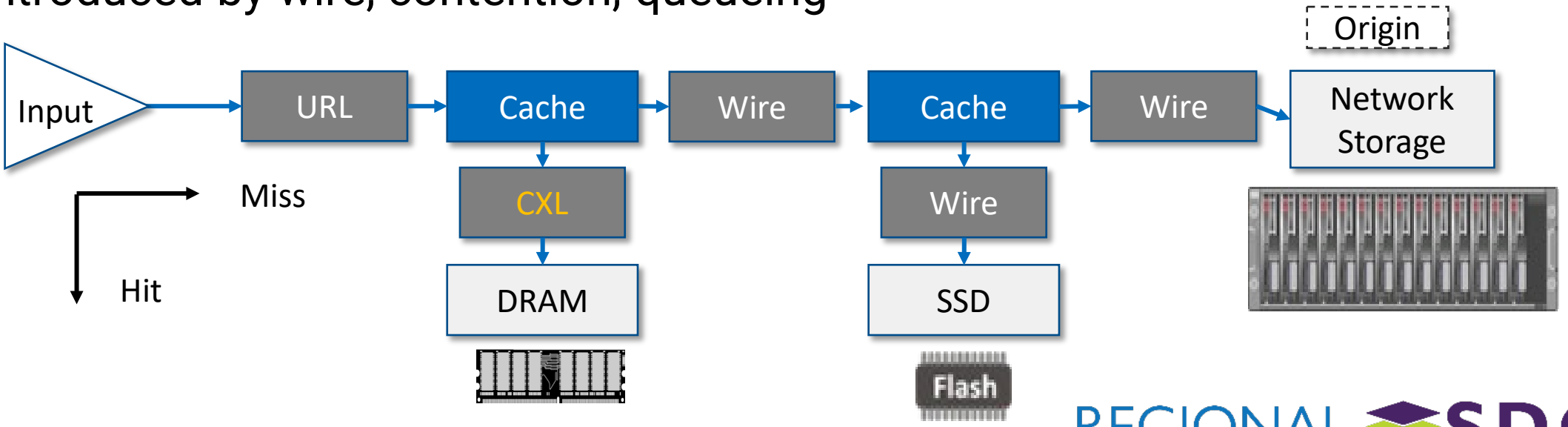
Memory bus, disk controller, network

Can multiplex and change form of IO request

Even type of wire can be variable

- Type of memory bus
- Hops in network topology

Delays introduced by wire, contention, queueing



Wire component

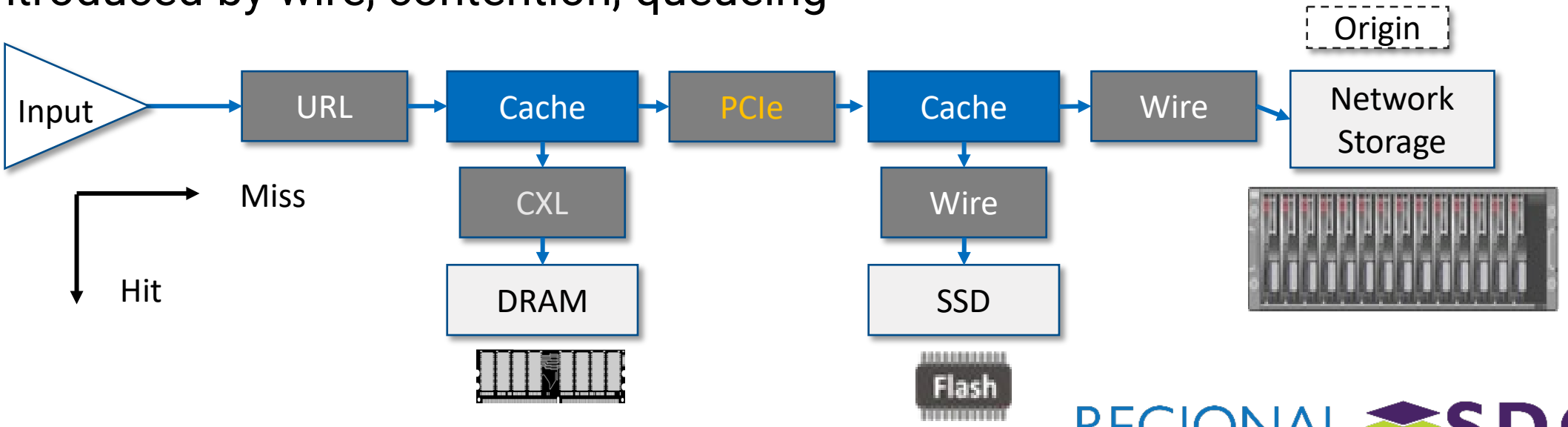
Memory bus, disk controller, network

Can multiplex and change form of IO request

Even type of wire can be variable

- Type of memory bus
- Hops in network topology

Delays introduced by wire, contention, queueing



Wire component

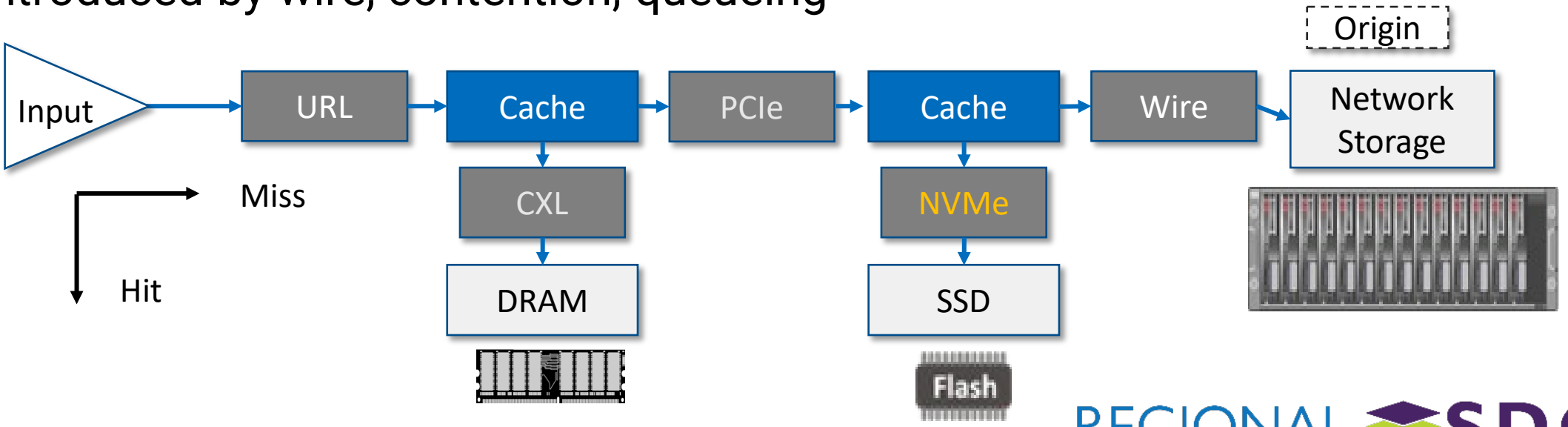
Memory bus, disk controller, network

Can multiplex and change form of IO request

Even type of wire can be variable

- Type of memory bus
- Hops in network topology

Delays introduced by wire, contention, queueing



Wire component

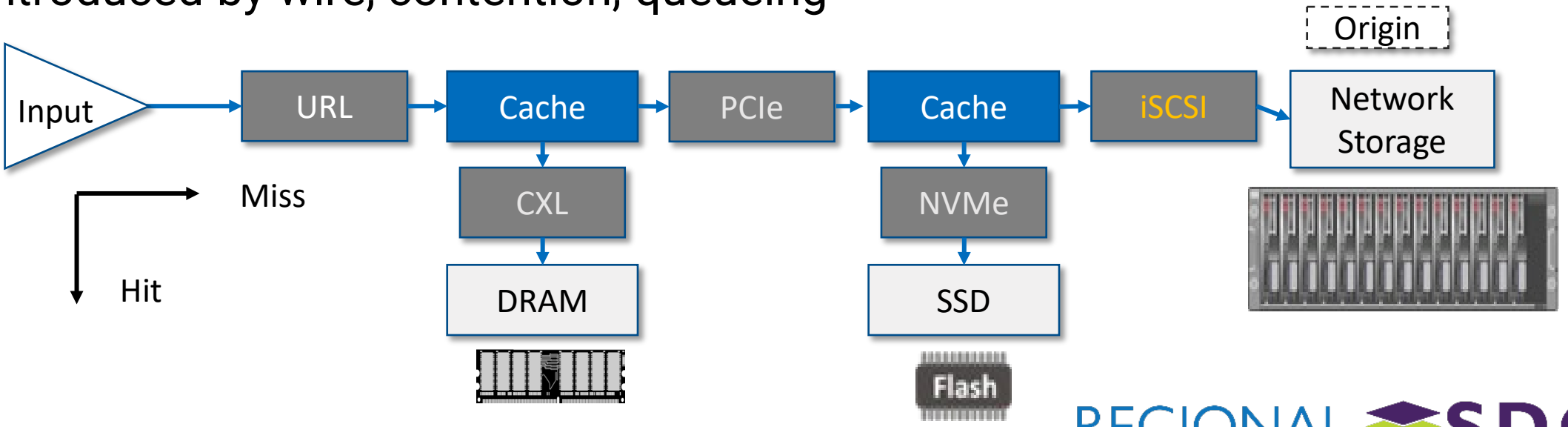
Memory bus, disk controller, network

Can multiplex and change form of IO request

Even type of wire can be variable

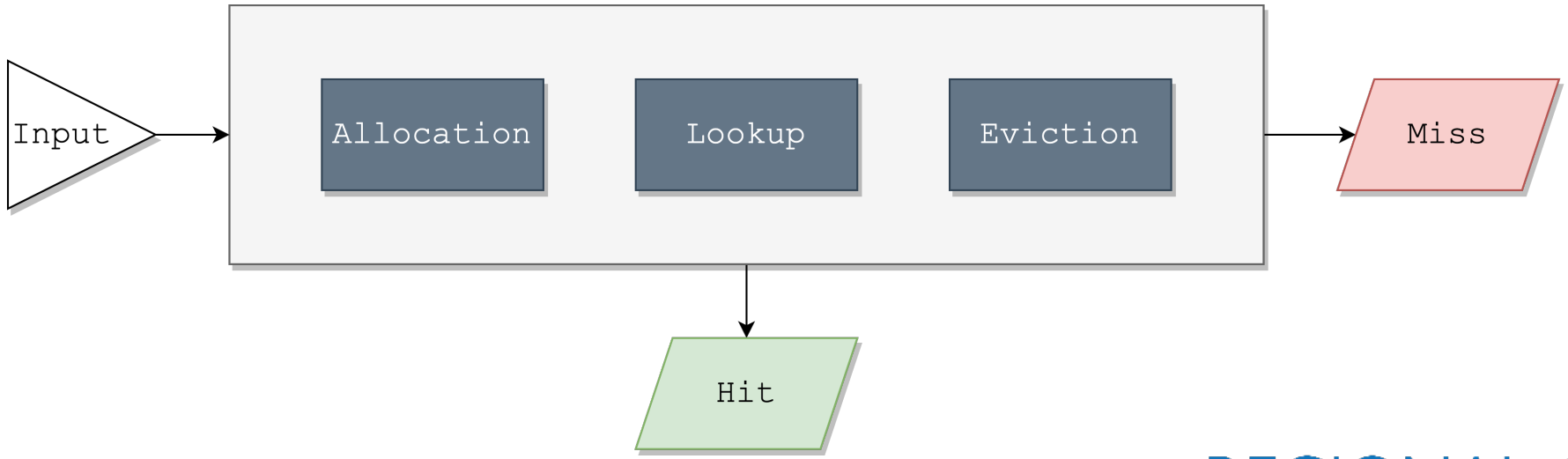
- Type of memory bus
- Hops in network topology

Delays introduced by wire, contention, queueing



Cache component

- Easily build basics like lookups, allocation, and eviction
- One (or more) hit path
- One (or more) miss path
- Many choices for variability



Workloads matter

No synthetic workloads

Content delivery

Learning and inference

Application storage IO

Simulation workloads

REGIONAL SDC 24

Workloads matter

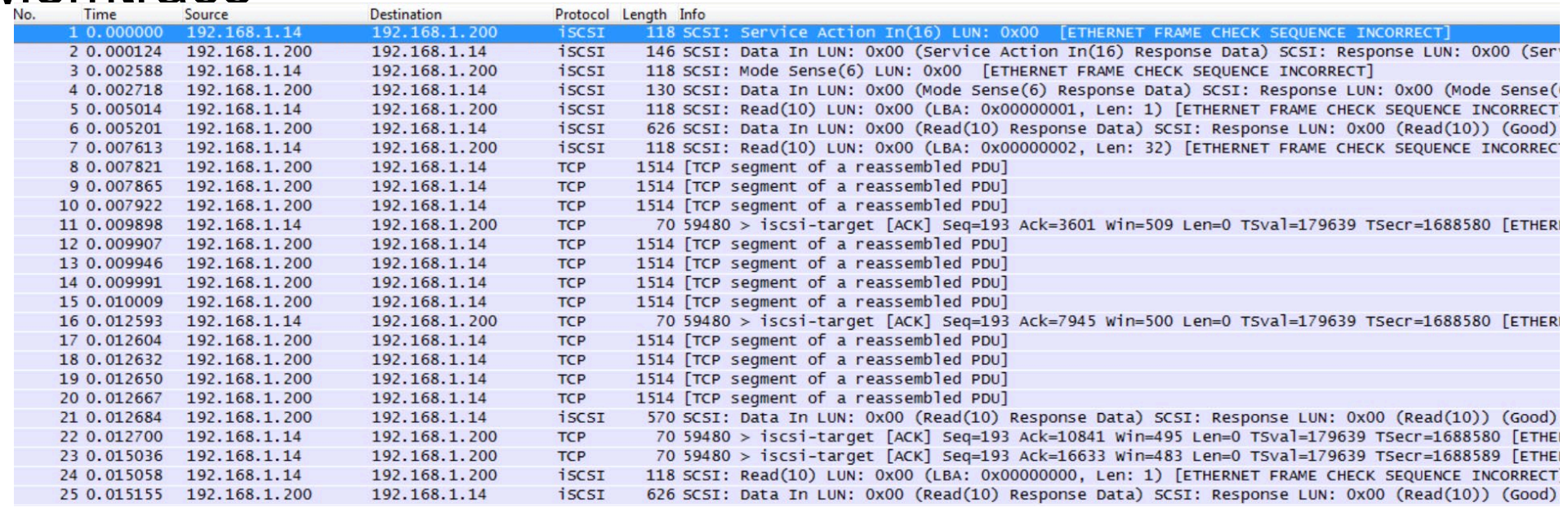
Flexible sources

VSCSI

Ethernet

HTTP

Memtrace

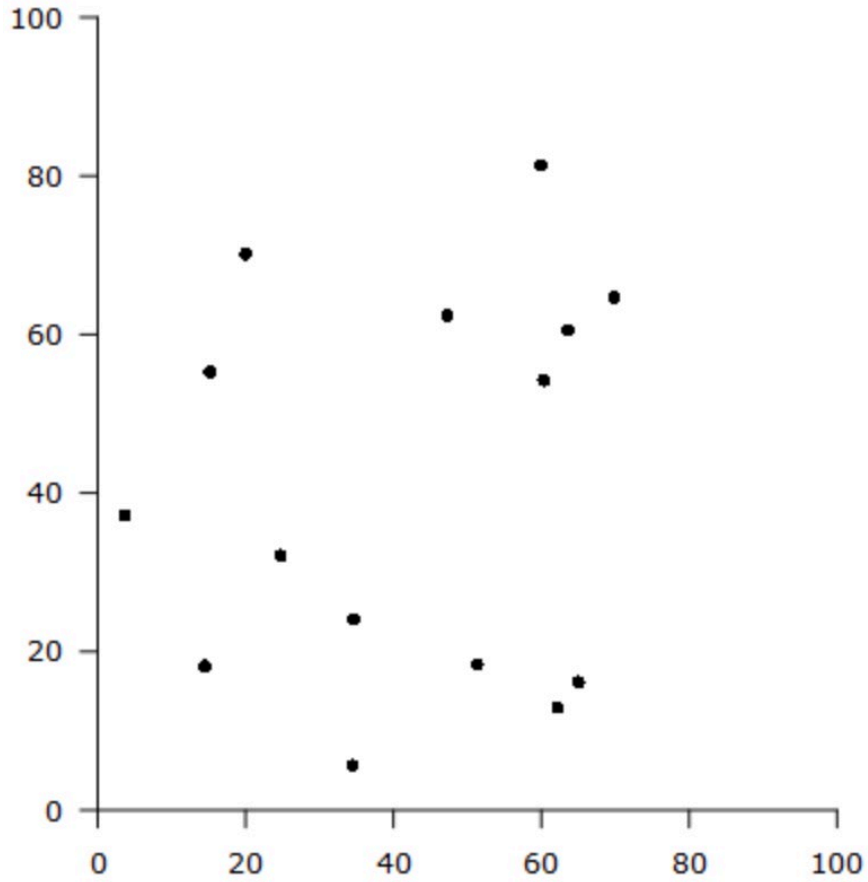


No.	Time	Source	Destination	Protocol	Length	Info
1	0.000000	192.168.1.14	192.168.1.200	iSCSI	118	SCSI: Service Action In(16) LUN: 0x00 [ETHERNET FRAME CHECK SEQUENCE INCORRECT]
2	0.000124	192.168.1.200	192.168.1.14	iSCSI	146	SCSI: Data In LUN: 0x00 (Service Action In(16) Response Data) SCSI: Response LUN: 0x00 (Ser
3	0.002588	192.168.1.14	192.168.1.200	iSCSI	118	SCSI: Mode Sense(6) LUN: 0x00 [ETHERNET FRAME CHECK SEQUENCE INCORRECT]
4	0.002718	192.168.1.200	192.168.1.14	iSCSI	130	SCSI: Data In LUN: 0x00 (Mode Sense(6) Response Data) SCSI: Response LUN: 0x00 (Mode Sense(
5	0.005014	192.168.1.14	192.168.1.200	iSCSI	118	SCSI: Read(10) LUN: 0x00 (LBA: 0x00000001, Len: 1) [ETHERNET FRAME CHECK SEQUENCE INCORRECT
6	0.005201	192.168.1.200	192.168.1.14	iSCSI	626	SCSI: Data In LUN: 0x00 (Read(10) Response Data) SCSI: Response LUN: 0x00 (Read(10)) (Good)
7	0.007613	192.168.1.14	192.168.1.200	iSCSI	118	SCSI: Read(10) LUN: 0x00 (LBA: 0x00000002, Len: 32) [ETHERNET FRAME CHECK SEQUENCE INCORREC
8	0.007821	192.168.1.200	192.168.1.14	TCP	1514	[TCP segment of a reassembled PDU]
9	0.007865	192.168.1.200	192.168.1.14	TCP	1514	[TCP segment of a reassembled PDU]
10	0.007922	192.168.1.200	192.168.1.14	TCP	1514	[TCP segment of a reassembled PDU]
11	0.009898	192.168.1.14	192.168.1.200	TCP	70	59480 > iscsi-target [ACK] Seq=193 Ack=3601 Win=509 Len=0 TSval=179639 TSecr=1688580 [ETHER
12	0.009907	192.168.1.200	192.168.1.14	TCP	1514	[TCP segment of a reassembled PDU]
13	0.009946	192.168.1.200	192.168.1.14	TCP	1514	[TCP segment of a reassembled PDU]
14	0.009991	192.168.1.200	192.168.1.14	TCP	1514	[TCP segment of a reassembled PDU]
15	0.010009	192.168.1.200	192.168.1.14	TCP	1514	[TCP segment of a reassembled PDU]
16	0.012593	192.168.1.14	192.168.1.200	TCP	70	59480 > iscsi-target [ACK] Seq=193 Ack=7945 Win=500 Len=0 TSval=179639 TSecr=1688580 [ETHER
17	0.012604	192.168.1.200	192.168.1.14	TCP	1514	[TCP segment of a reassembled PDU]
18	0.012632	192.168.1.200	192.168.1.14	TCP	1514	[TCP segment of a reassembled PDU]
19	0.012650	192.168.1.200	192.168.1.14	TCP	1514	[TCP segment of a reassembled PDU]
20	0.012667	192.168.1.200	192.168.1.14	TCP	1514	[TCP segment of a reassembled PDU]
21	0.012684	192.168.1.200	192.168.1.14	iSCSI	570	SCSI: Data In LUN: 0x00 (Read(10) Response Data) SCSI: Response LUN: 0x00 (Read(10)) (Good)
22	0.012700	192.168.1.14	192.168.1.200	TCP	70	59480 > iscsi-target [ACK] Seq=193 Ack=10841 Win=495 Len=0 TSval=179639 TSecr=1688580 [ETHE
23	0.015036	192.168.1.14	192.168.1.200	TCP	70	59480 > iscsi-target [ACK] Seq=193 Ack=16633 Win=483 Len=0 TSval=179639 TSecr=1688589 [ETHE
24	0.015058	192.168.1.14	192.168.1.200	iSCSI	118	SCSI: Read(10) LUN: 0x00 (LBA: 0x00000000, Len: 1) [ETHERNET FRAME CHECK SEQUENCE INCORRECT
25	0.015155	192.168.1.200	192.168.1.14	iSCSI	626	SCSI: Data In LUN: 0x00 (Read(10) Response Data) SCSI: Response LUN: 0x00 (Read(10)) (Good)

Use Cases

Objectives

Fixed costs of BOM

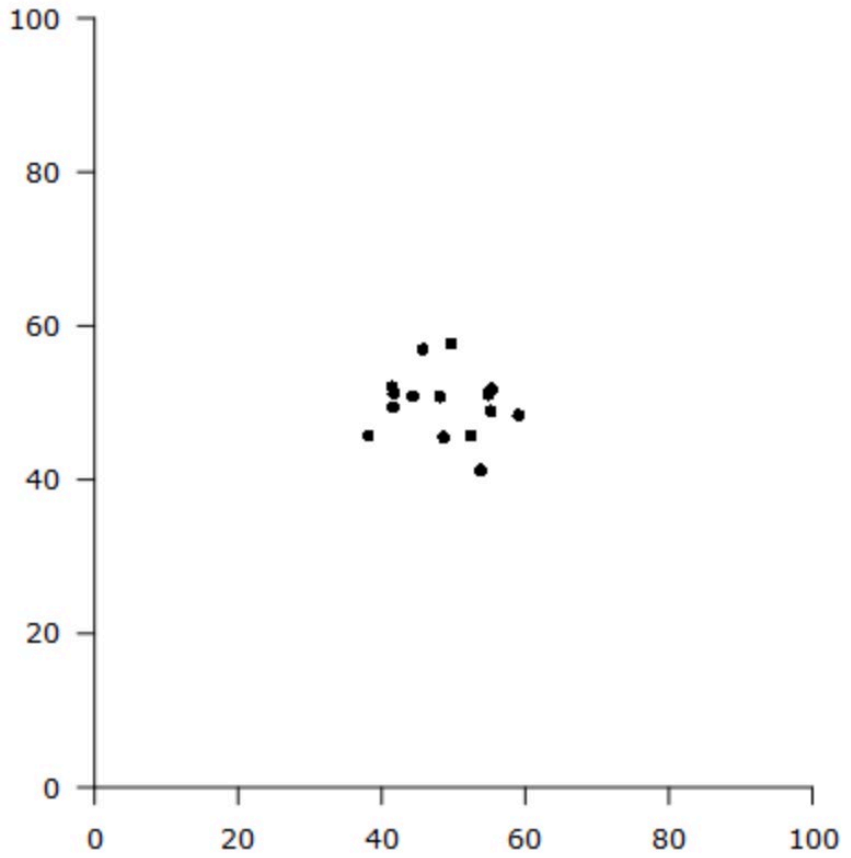


REGIONAL SDC 24

Objectives

Fixed costs of BOM

Re-use of existing datacenter/infrastructure

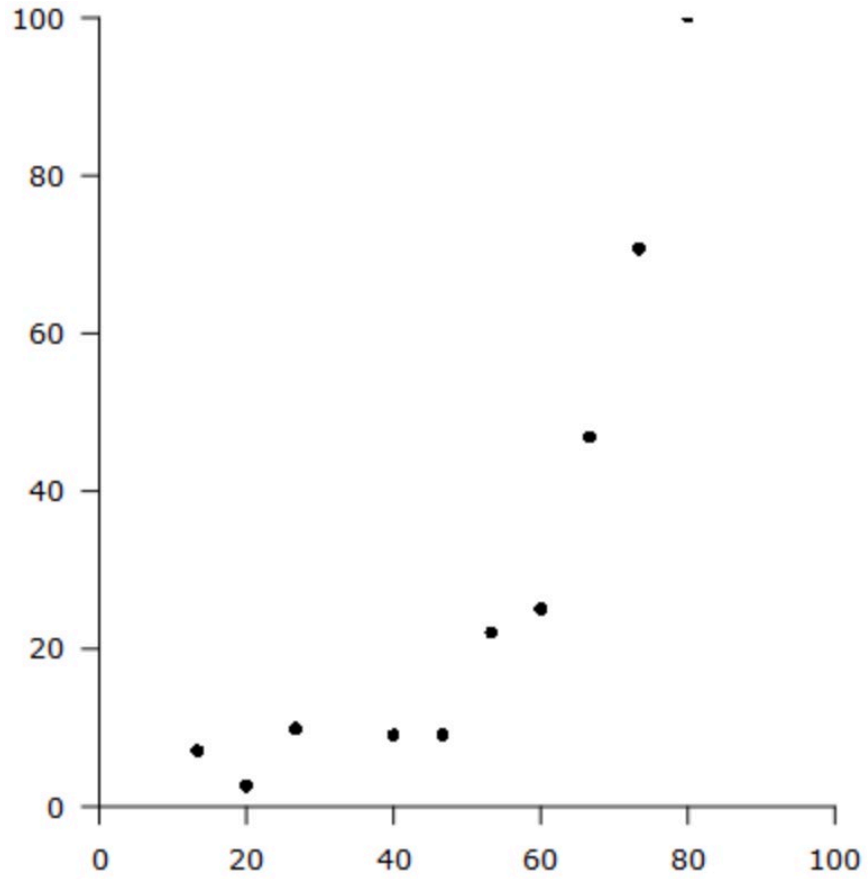


Objectives

Fixed costs of BOM

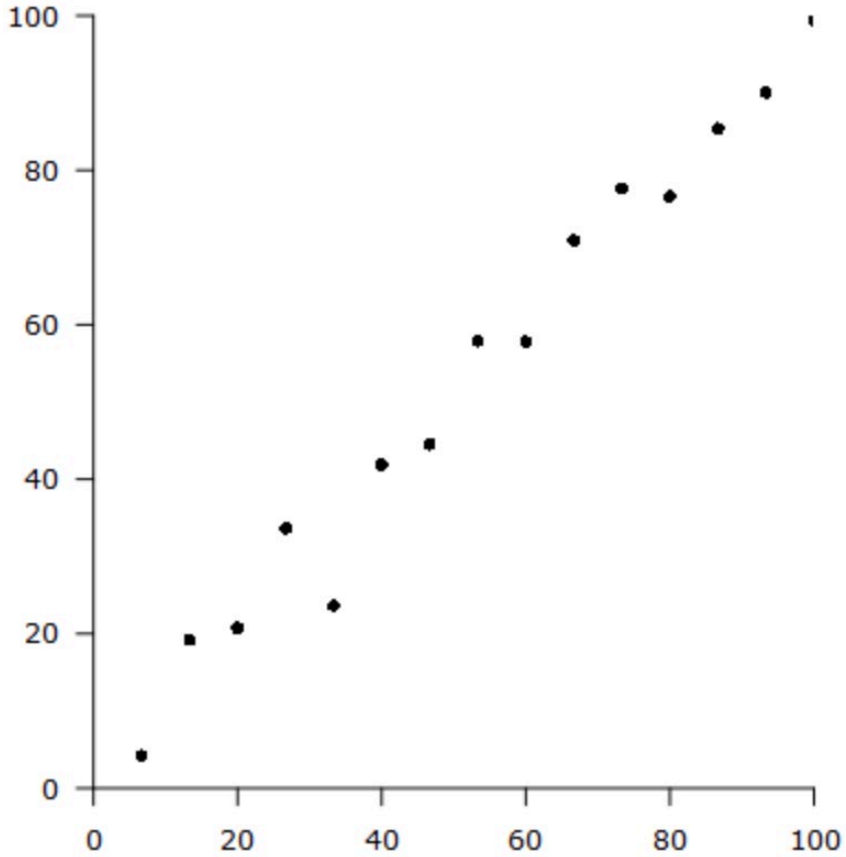
Re-use of existing datacenter/infrastructure

Identifying wasted resources



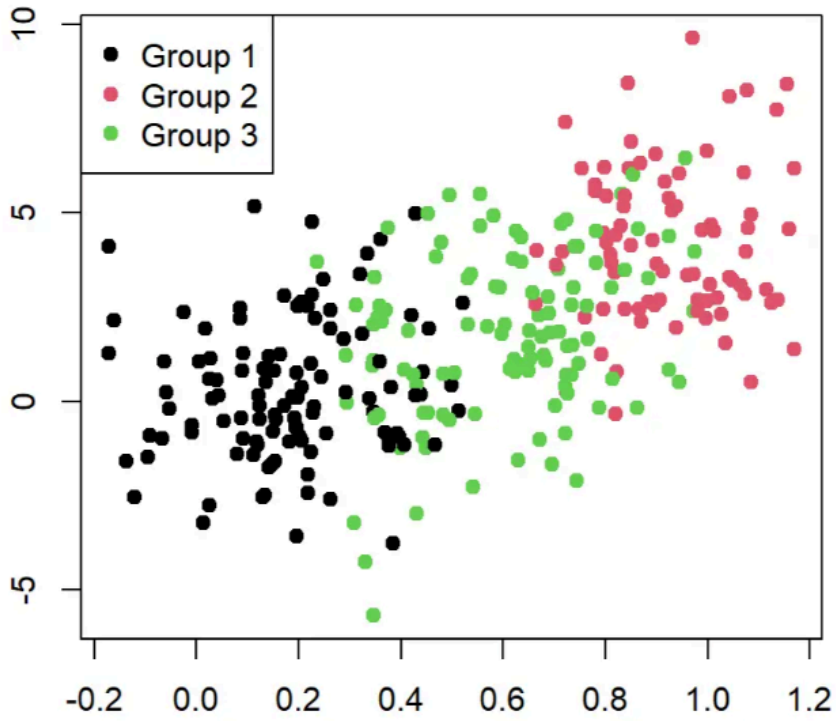
Objectives

- Fixed costs of BOM
- Re-use of existing datacenter/infrastructure
- Identifying wasted resources
- Best performance for a given workload



Objectives

- Fixed costs of BOM
- Re-use of existing datacenter/infrastructure
- Identifying wasted resources
- Best performance for a given workload
- Best cost or performance for a variety of workloads



Objectives

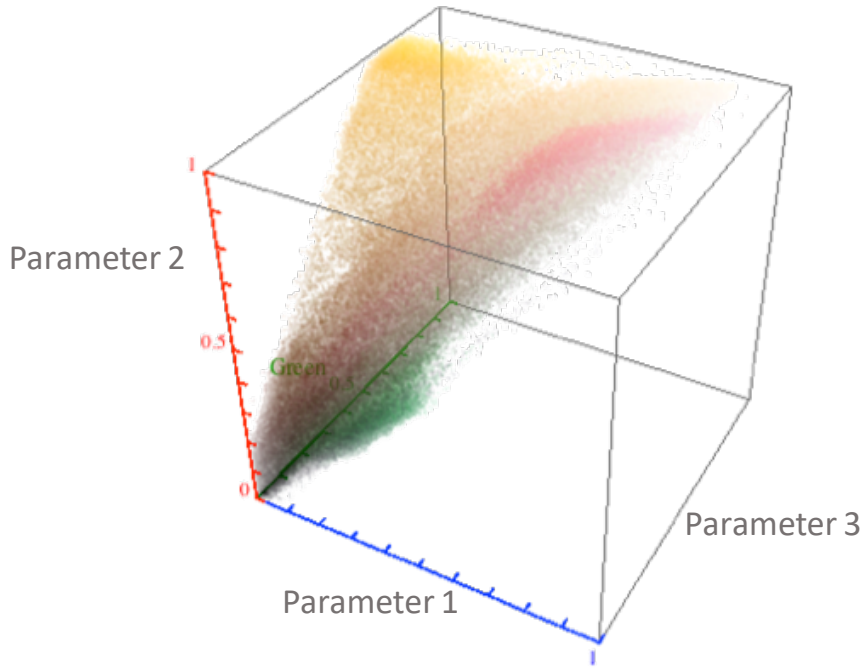
Fixed costs of BOM

Re-use of existing datacenter/infrastructure

Identifying wasted resources

Best performance for a given workload

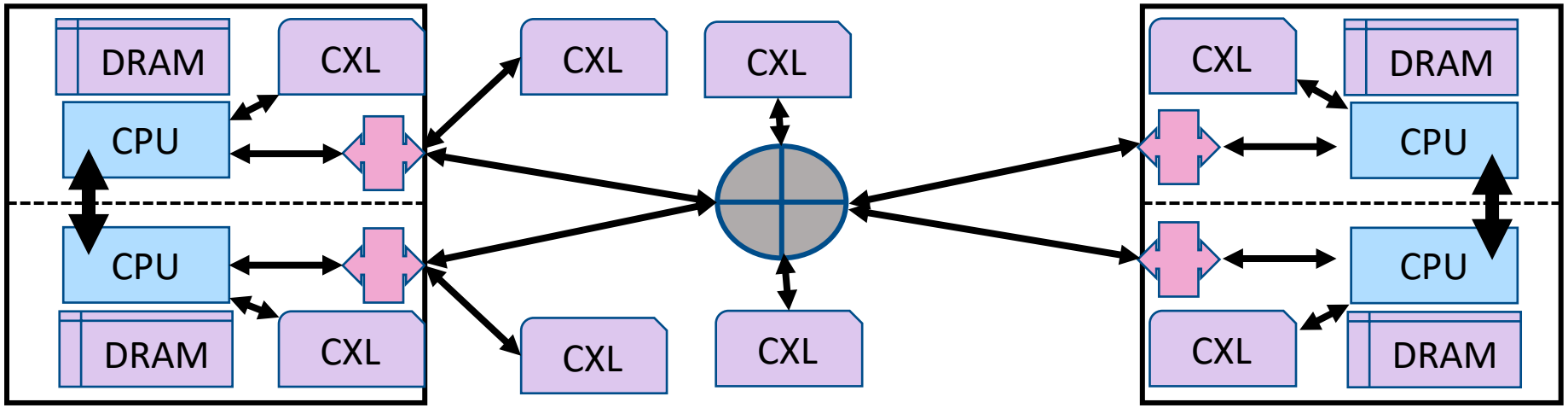
Best cost or performance for a variety of workloads



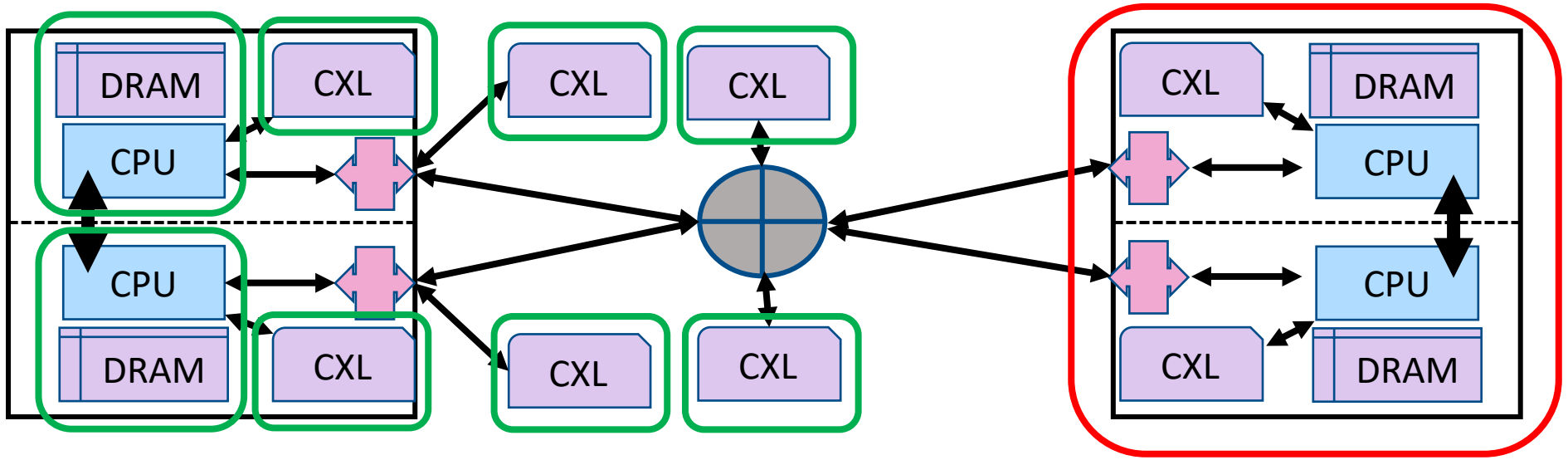
REGIONAL SDC 24

Example Use Cases

Use case: Memory tiering in CXL hosts

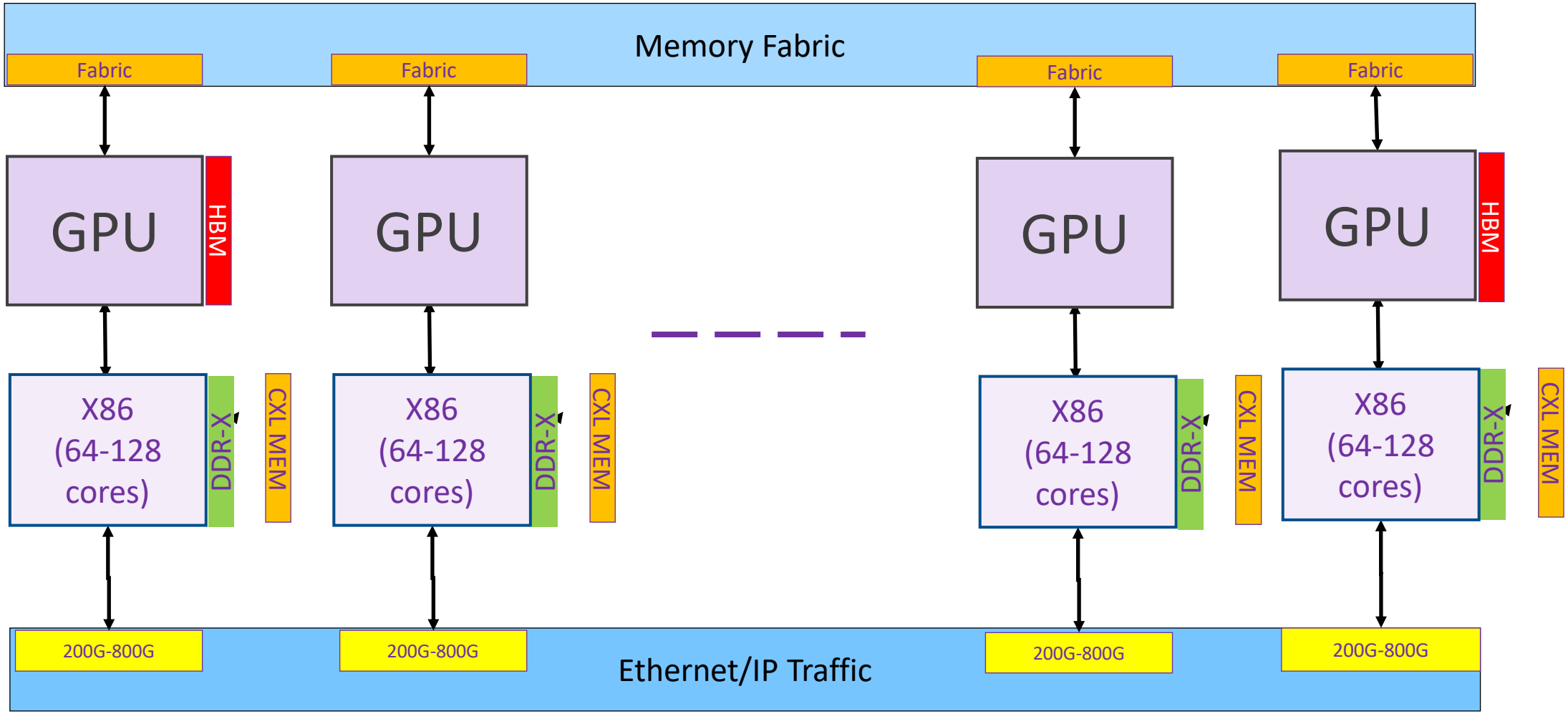


Use case: Memory tiering in CXL hosts

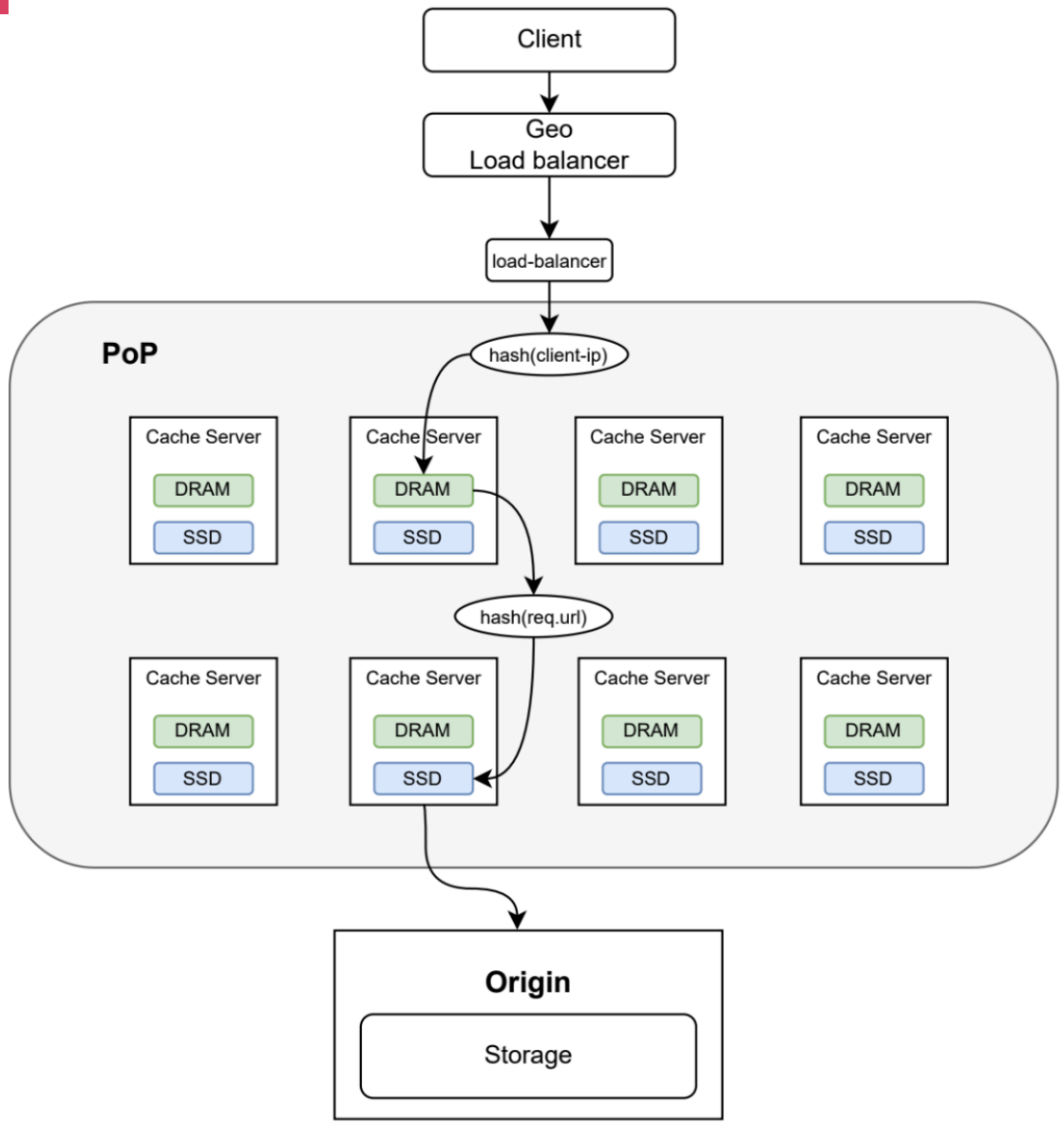


7 different NUMA nodes plus contention

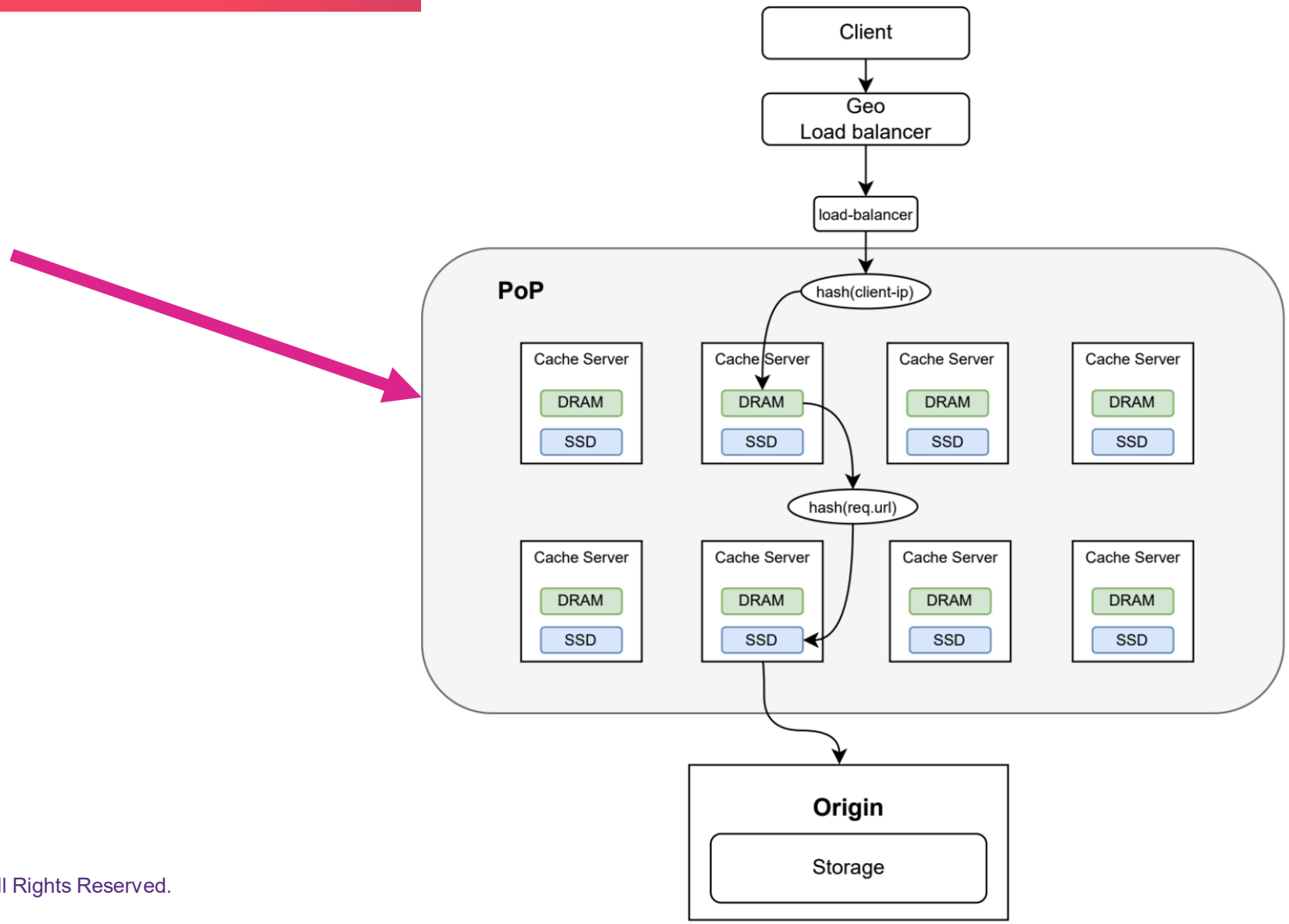
Use case: Datacenter/Hyperscaler cluster



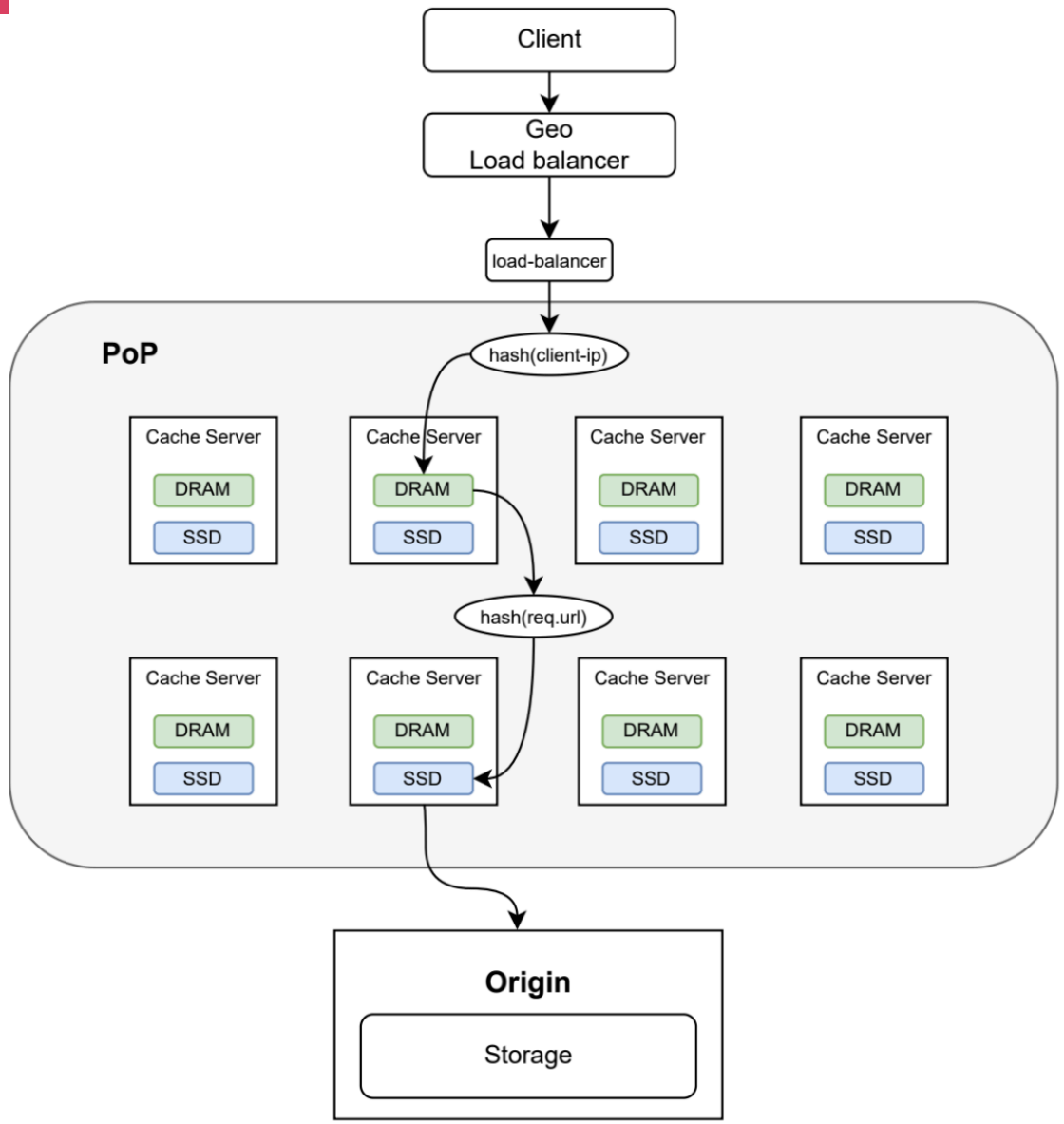
Use case: Multilayer Content Delivery



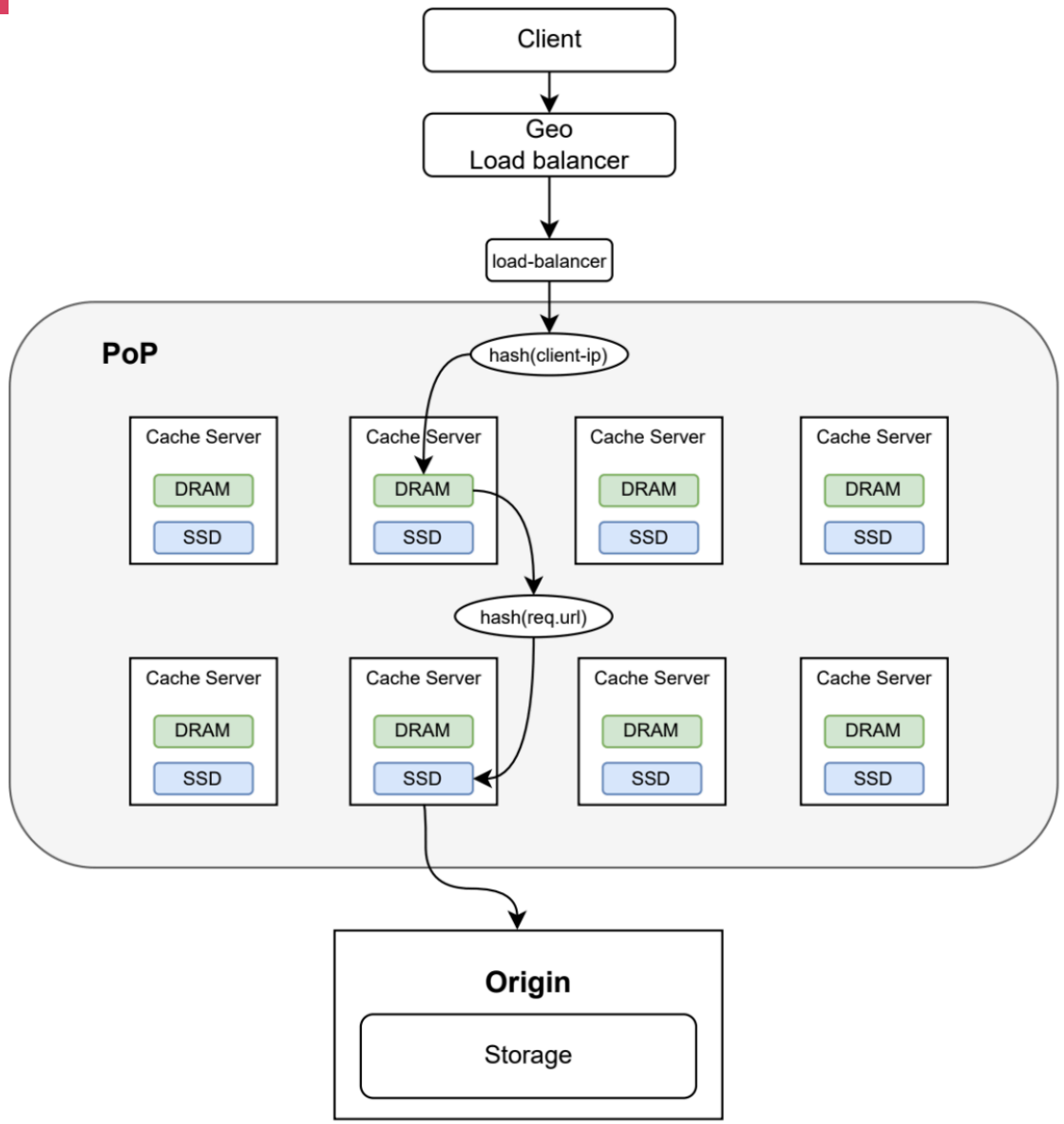
Use case: Multilayer Content Delivery



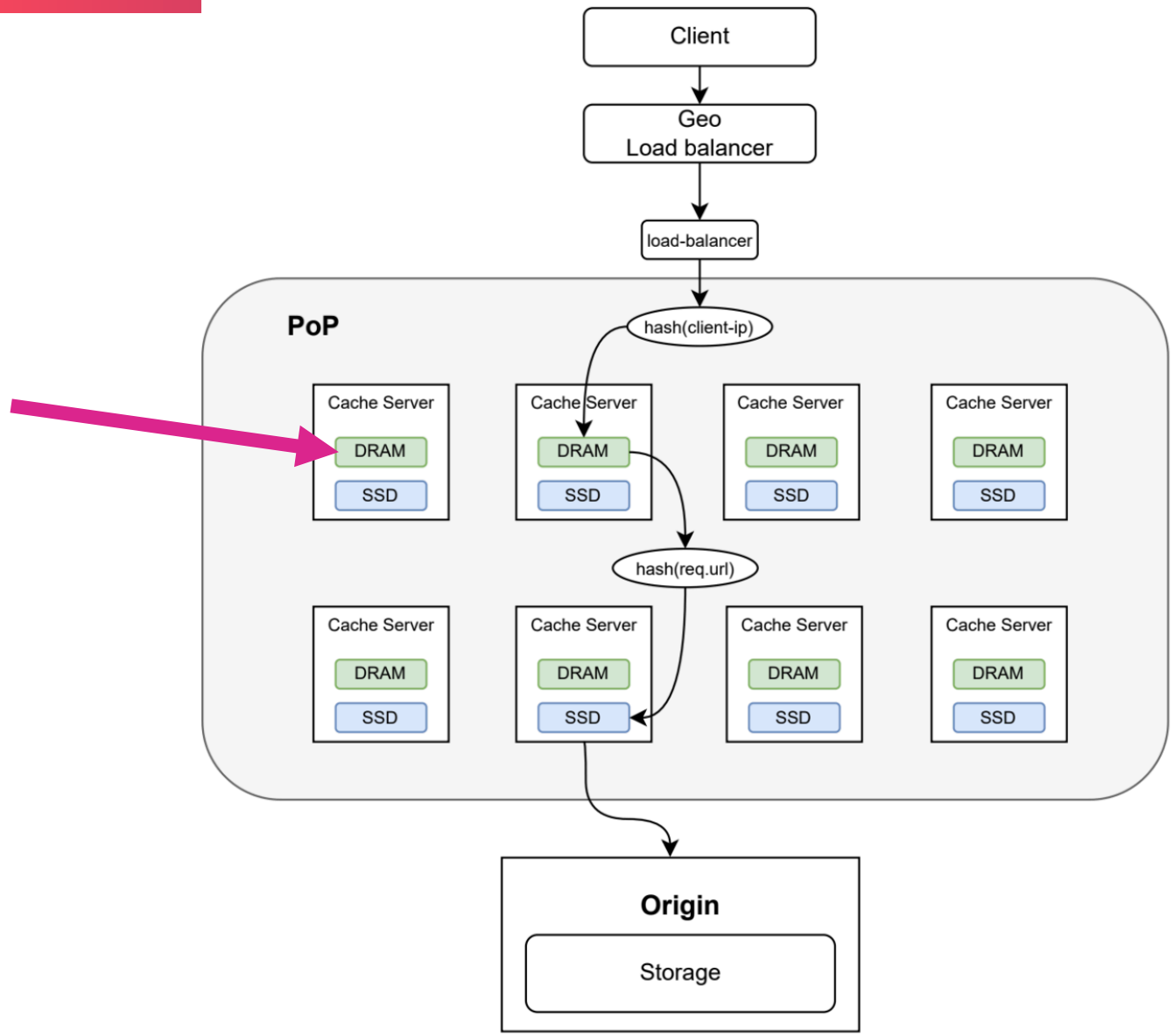
Use case: Multilayer Content Delivery



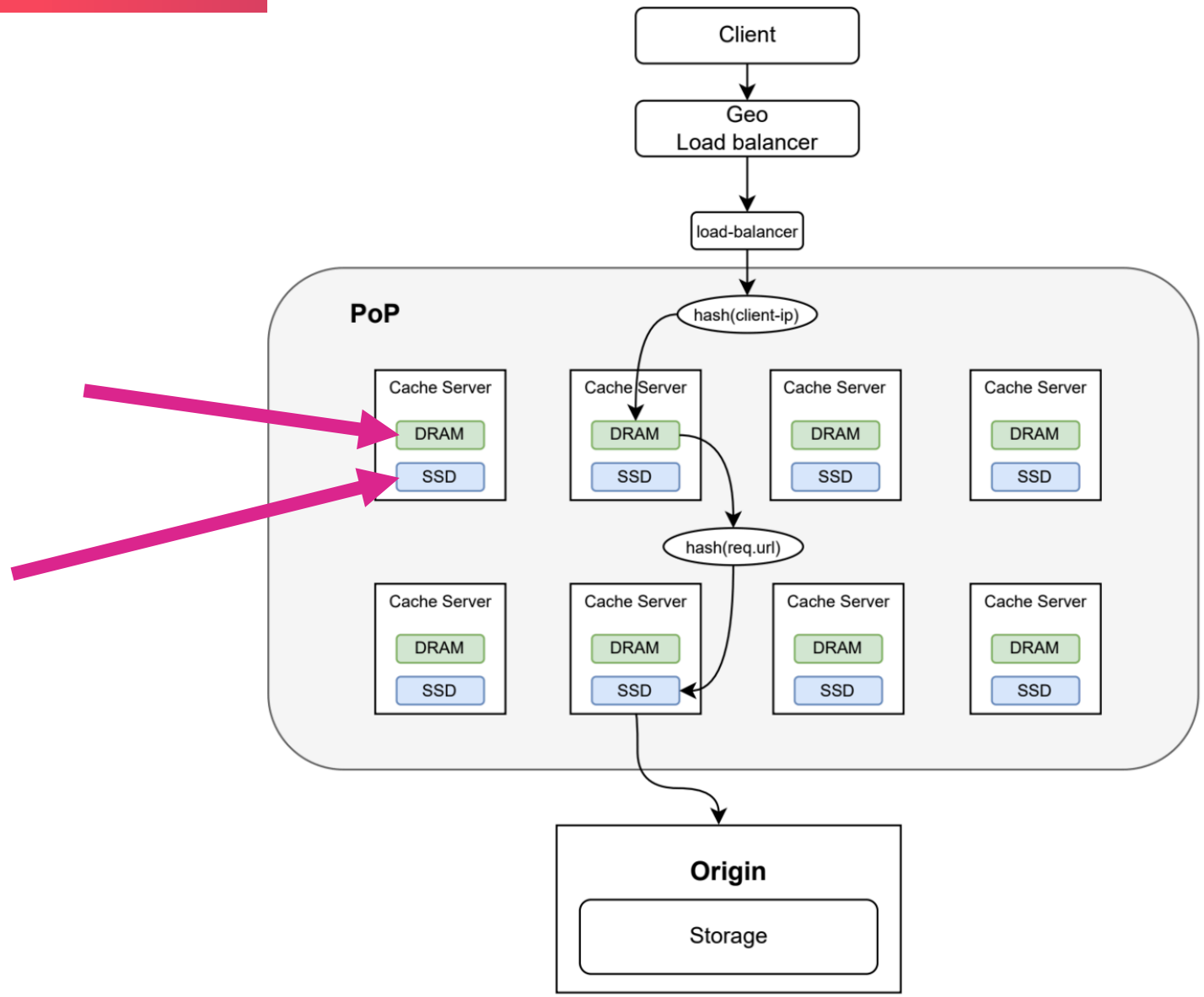
Use case: Multilayer Content Delivery



Use case: Multilayer Content Delivery

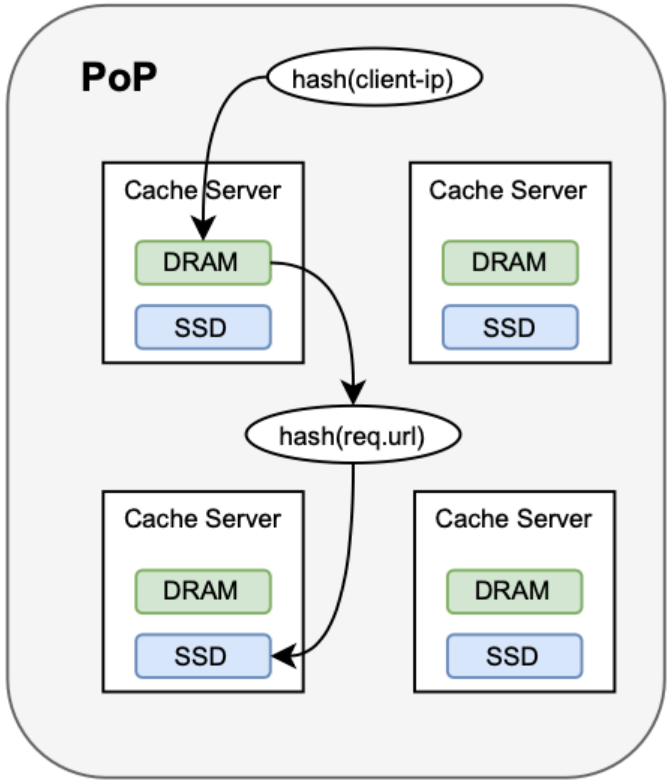


Use case: Multilayer Content Delivery

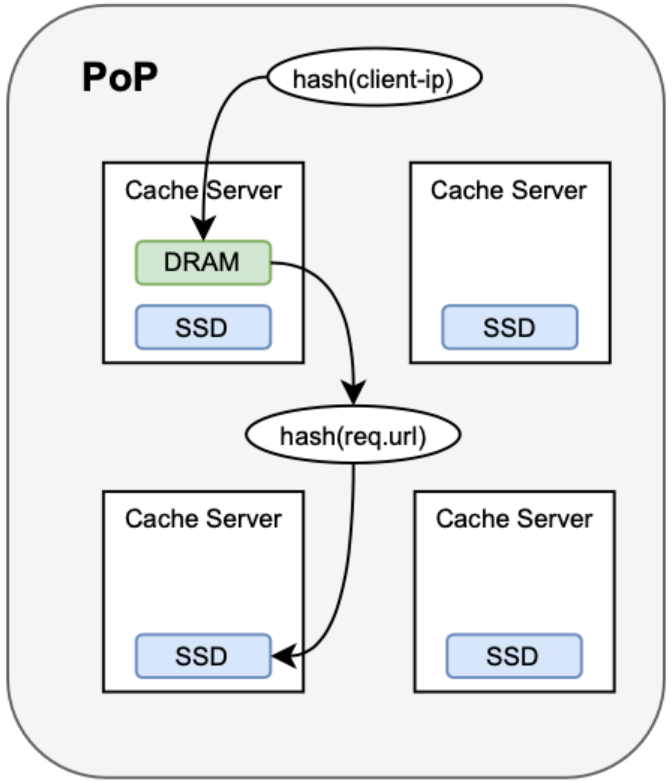


Use case: Multilayer Content Delivery

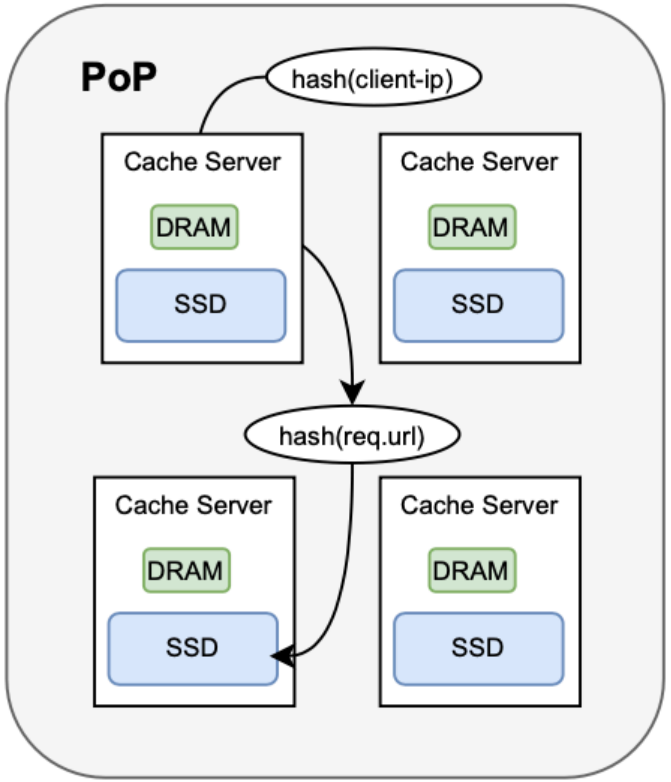
1. Varying the number of cache servers



2. Varying the number of DRAM and SSD



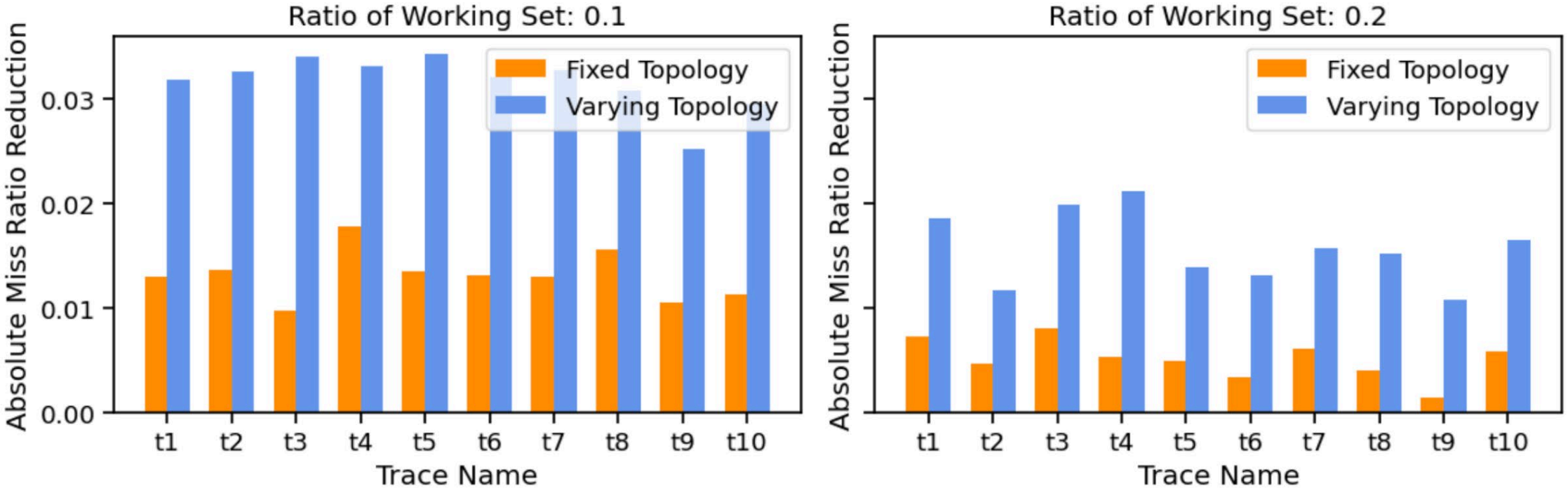
3. Varying the capacity of DRAM and SSD



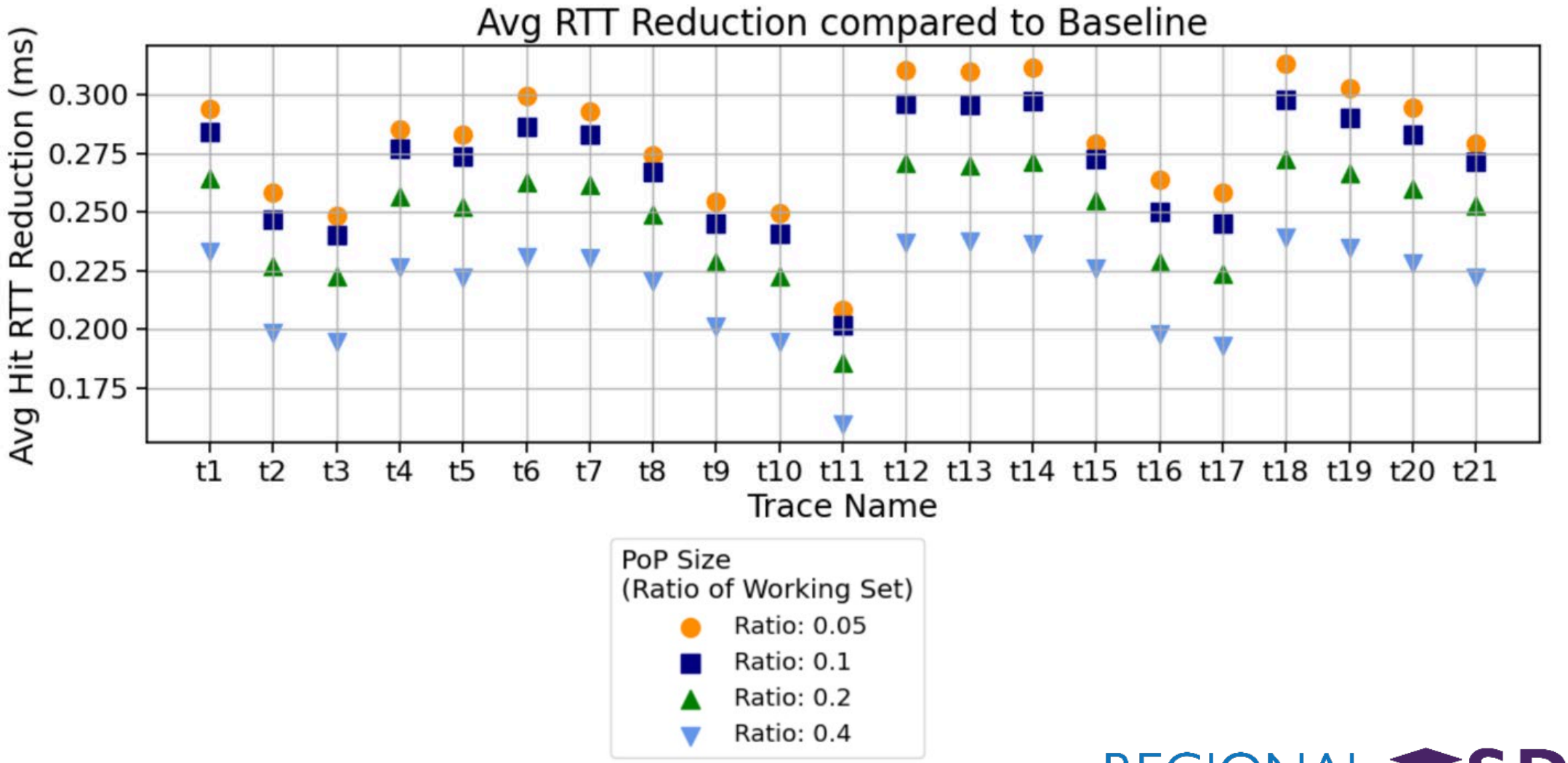
REGIONAL SDC 24

Understandable Results

CDN results



CDN results



RESULTS WITH MAGNITION

PROVEN IN MARKET TODAY

As an example, a current customer has achieved the following measurable outcomes with Magnition:

Experiments **per day per engineer**:

- Without Magnition**: 2**
- With Magnition: **50,000+**

Parameter variations tested **before prod release**:

- Without Magnition**: 50**
- With Magnition: **1,000,000+**

Workload performance improvement using our products to find **optimal out-of-the-box settings**: **10-50%+**

REGIONAL SDC 24

ABOUT MAGNITION

STORAGE PERFORMANCE, REINVENTED

World's First Real-Time Data Placement Optimization

Patented technology is a first for the industry.

Proven At-Scale, with Production Workloads

Use customer traces to fully test diverse workloads in real-time.

Peer-Reviewed and Published in Leading Journals

Multiple industry articles published and reviewed.

Award-Winning, Patented Technology

3-time award winner for innovative technology.

THANK YOU

Please take a moment to rate this session.

REGIONAL SDC 24

BY Developers FOR Developers