Revisiting Memory Errors in Large-Scale Production Data Centers

Analysis and Modeling of New Trends from the Field

Justin Meza
Qiang Wu
Sanjeev Kumar
Onur Mutlu

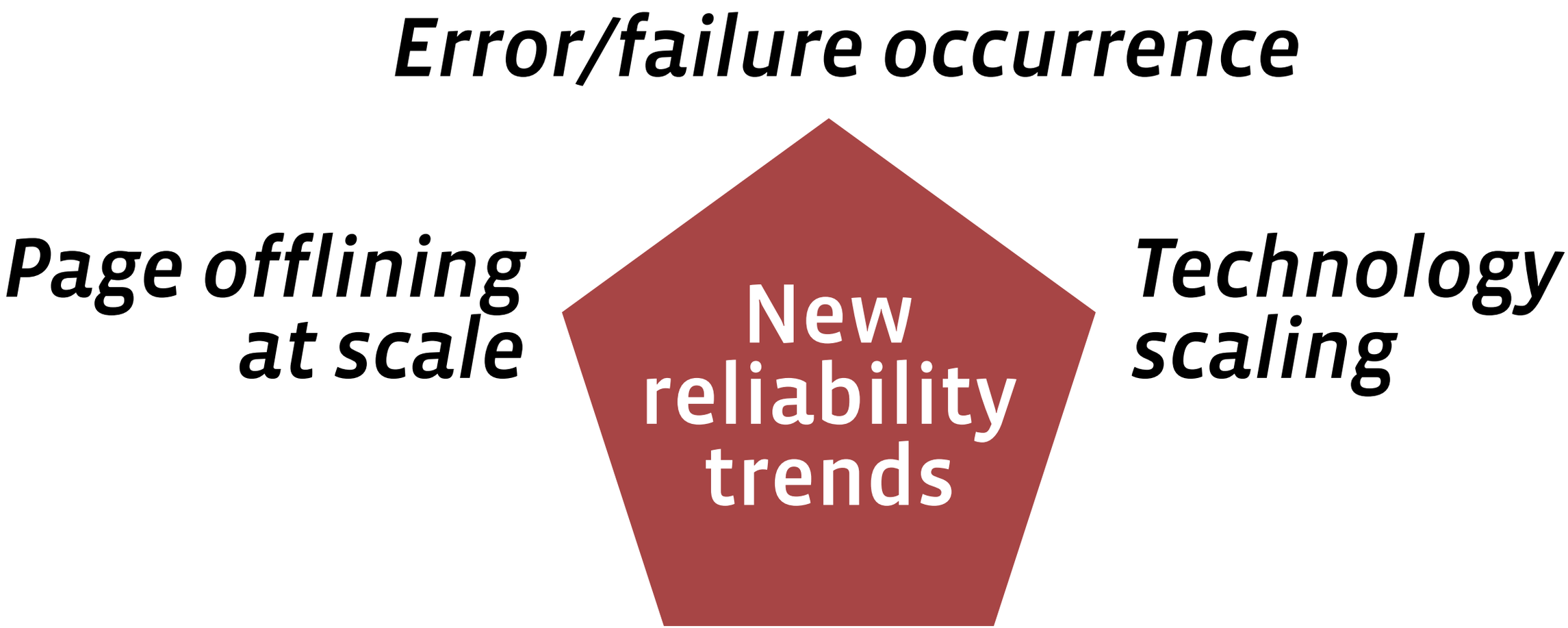
facebook
Carnegie Mellon University

Overview

Study of DRAM reliability:

- on ***modern*** devices and workloads
- at a ***large scale*** in the field

Overview


Error/failure occurrence
Page offlining at scale
New reliability trends
Technology scaling
Modeling errors
Architecture & workload

Overview

Error/failure occurrence

Errors follow a ***power-law distribution*** and a large number of errors occur due to ***sockets/channels***

Modeling errors

Architecture & workload

Overview

Error/failure occurrence

We find that ***newer*** cell fabrication technologies have ***higher failure rates***

Technology scaling

trends

Modeling errors

Architecture & workload

Overview

Error/failure occurrence

Chips per DIMM, ***transfer width***, and ***workload type*** (not necessarily CPU/ memory utilization) affect reliability

Page c *ogy*

trends

Modeling errors

Architecture & workload

Overview

Error/failure occurrence

We have made publicly available a ***statistical model*** for assessing server memory reliability

Page c *ogy*

trends

Modeling errors

Architecture & workload

Overview

Error/failure occurrence

Page offlining at scale

First large-scale study of page offlining; real-world ***limitations*** of technique

trends

Modeling errors

Architecture & workload

Outline

- background and motivation
- server memory organization
- error collection/analysis methodology
- memory reliability trends
- summary

Background and motivation

DRAM errors are common

- examined extensively in prior work
  - charged particles, wear-out
  - variable retention time (next talk)
- error correcting codes
  - used to detect and correct errors
  - require additional storage overheads

Our goal

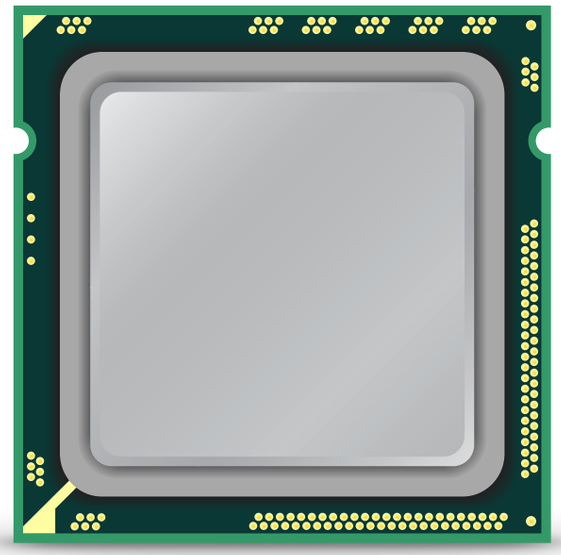
Strengthen understanding of DRAM reliability by studying:

- new trends in DRAM errors
  - modern devices and workloads
- at a large scale
  - billions of device-days, across 14 months

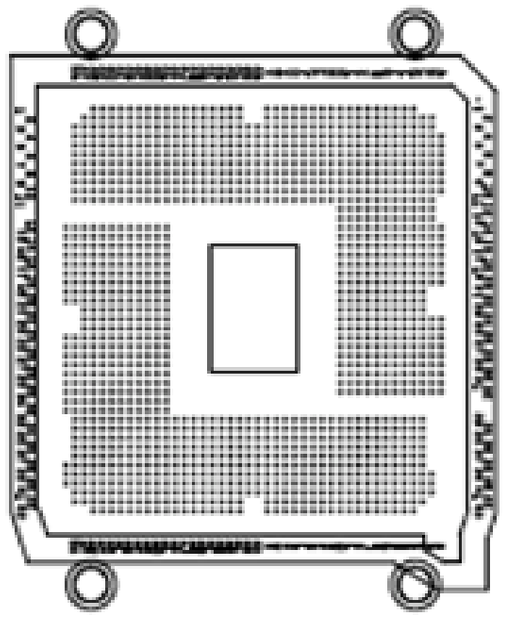
Our main contributions

- identified ***new*** DRAM failure trends
- developed a ***model*** for DRAM errors
- evaluated ***page offlining at scale***

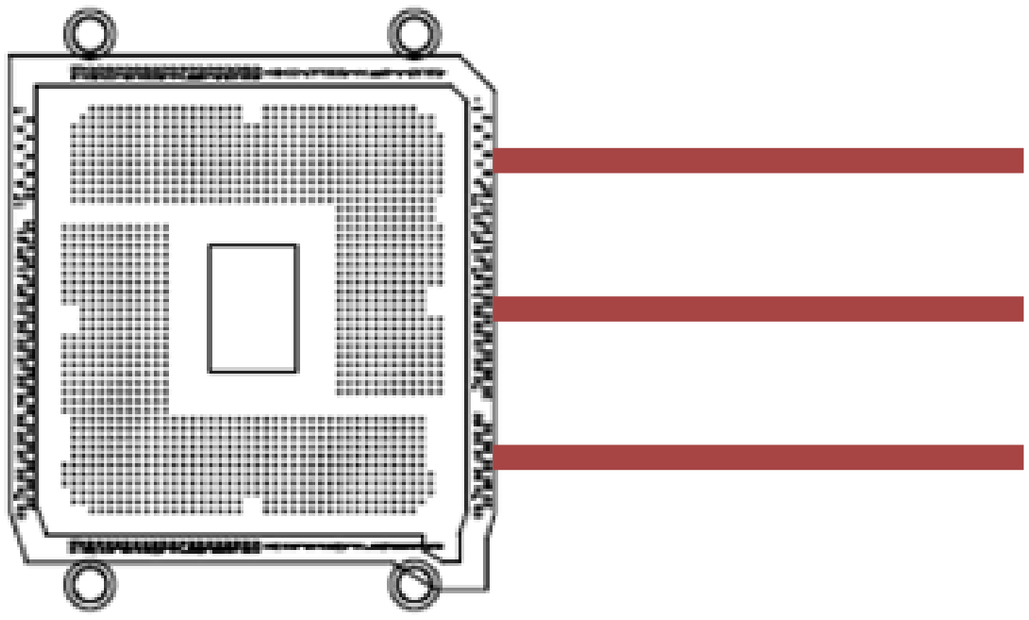
Server memory organization

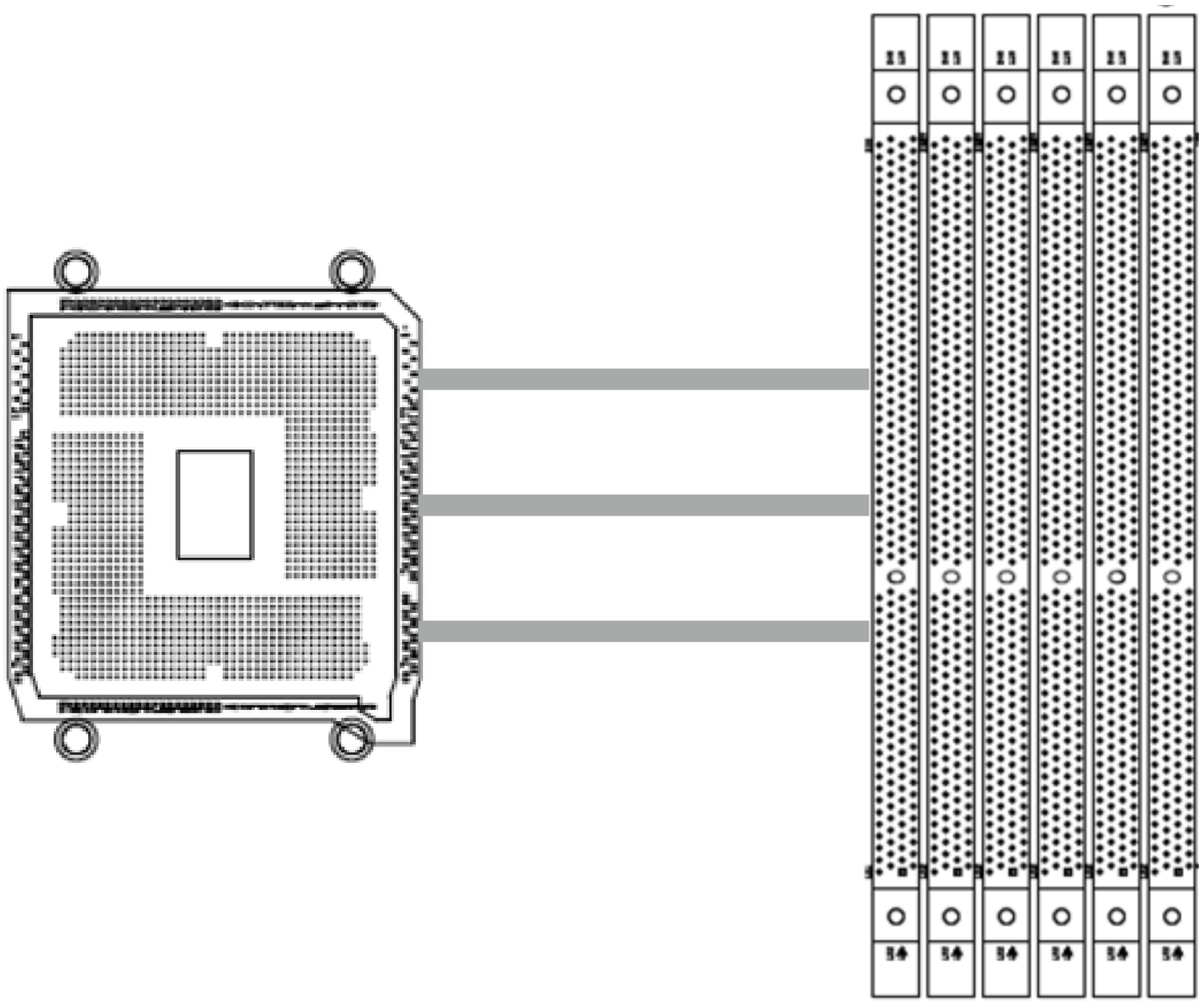


Socket



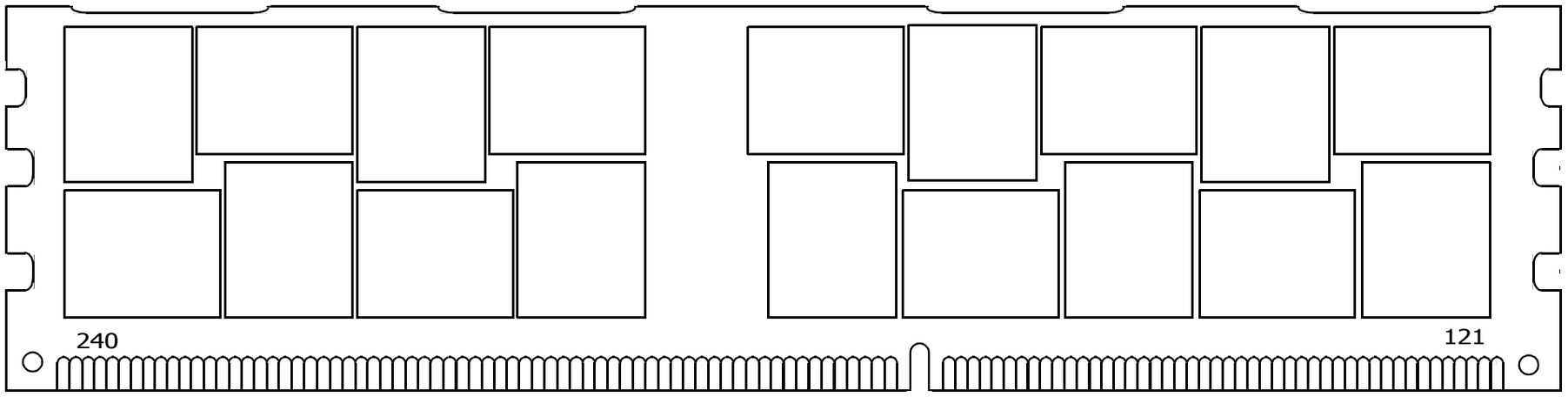
Memory channels

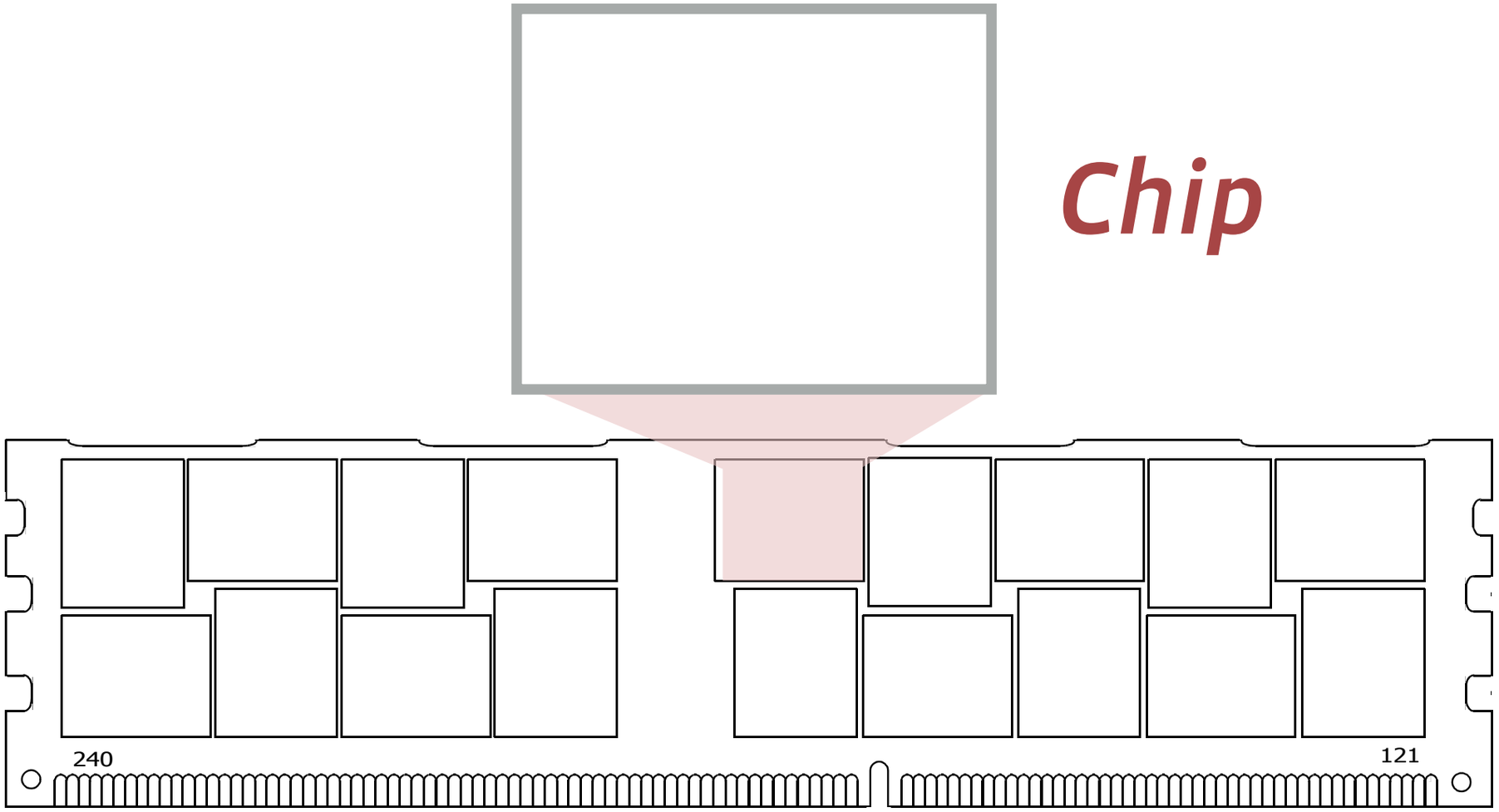




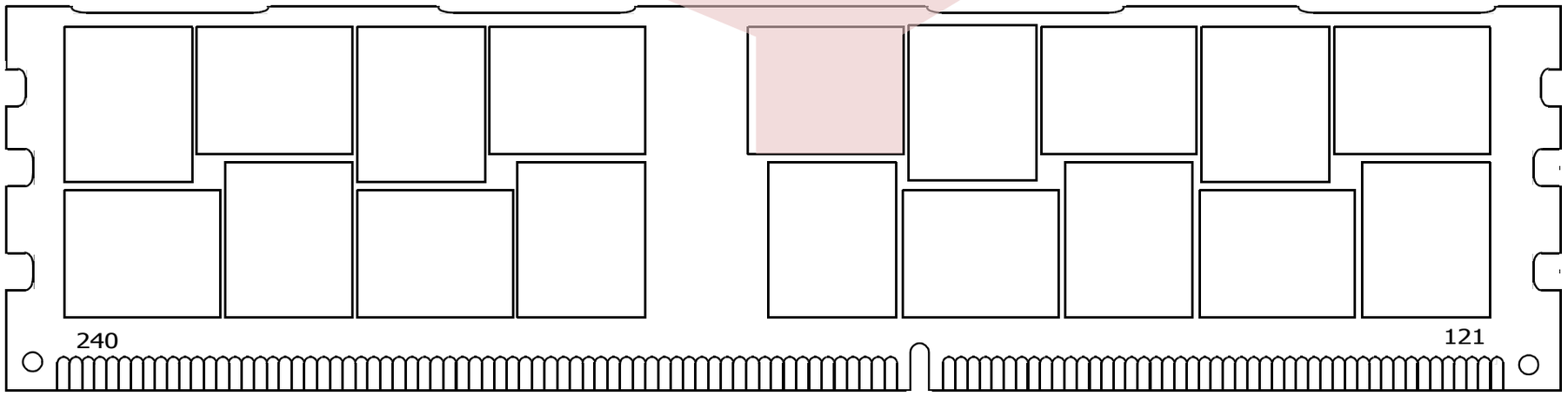
DIMM slots

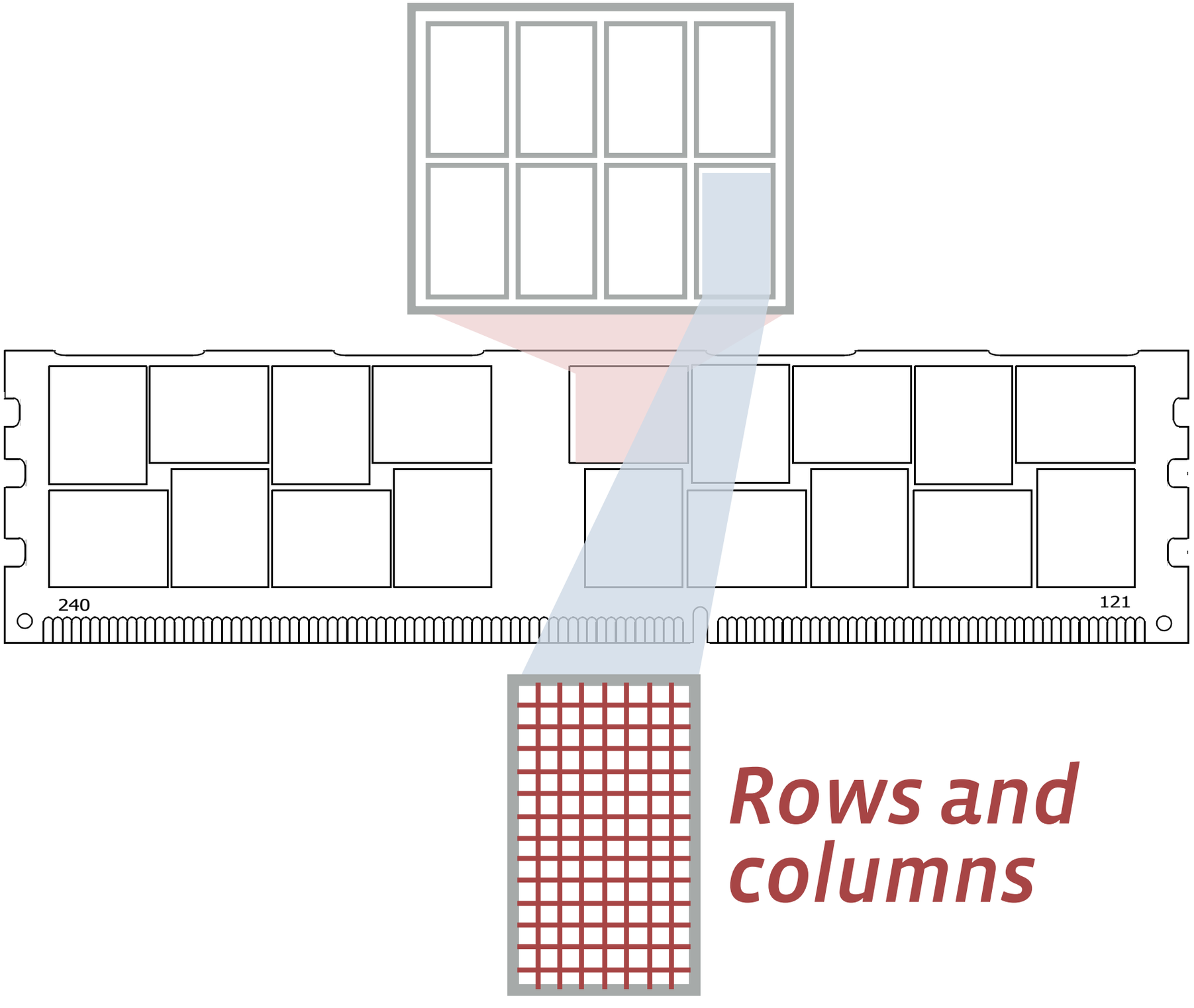
DIMM



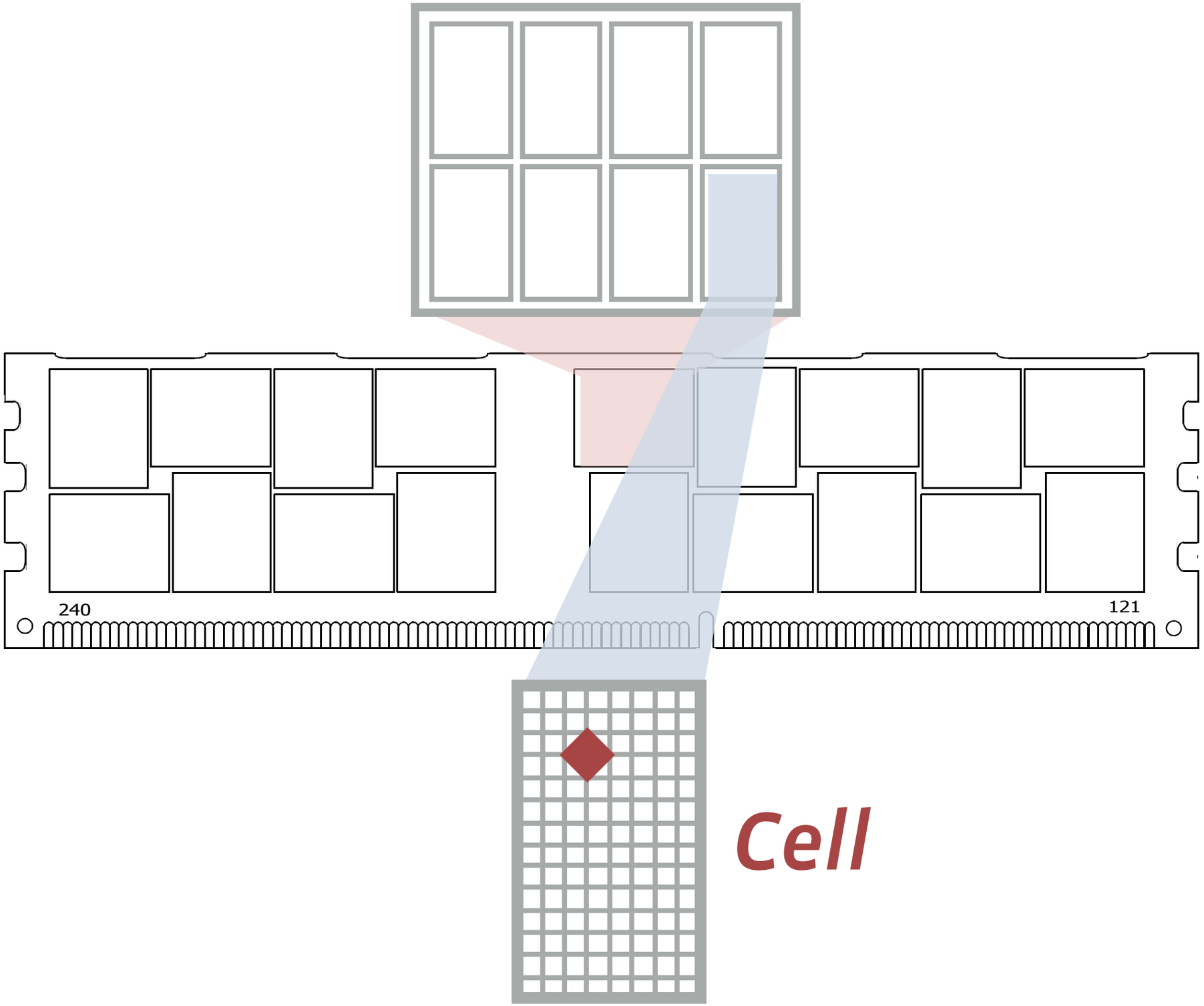

Chip
240
121

Banks

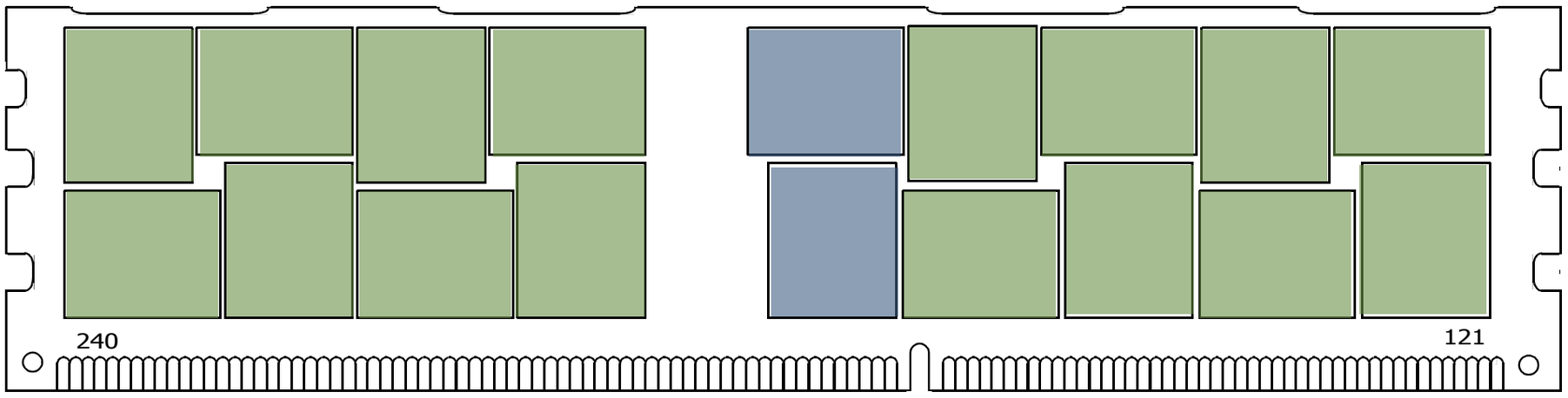




Rows and columns


240
121
Cell

User data



ECC metadata

additional 12.5% overhead

Reliability events

Fault

- the underlying cause of an error
  - DRAM cell unreliably stores charge

Error

- the manifestation of a fault
- ***permanent:*** every time
- ***transient:*** only some of the time

Error collection/ analysis methodology

DRAM error measurement

- measured every correctable error
  - across Facebook's fleet
  - for 14 months
    - metadata associated with each error
- parallelized Map-Reduce to process
- used R for further analysis

System characteristics

- 6 different system configurations
  - Web, Hadoop, Ingest, Database, Cache, Media
  - diverse CPU/memory/storage requirements
- ***modern*** DRAM devices
  - DDR3 communication protocol
    - (more aggressive clock frequencies)
  - diverse organizations (banks, ranks, …)
  - previously unexamined characteristics
    - density, # of chips, transfer width, workload

Memory reliability trends

Error/failure occurrence

Page offlining at scale

New reliability trends

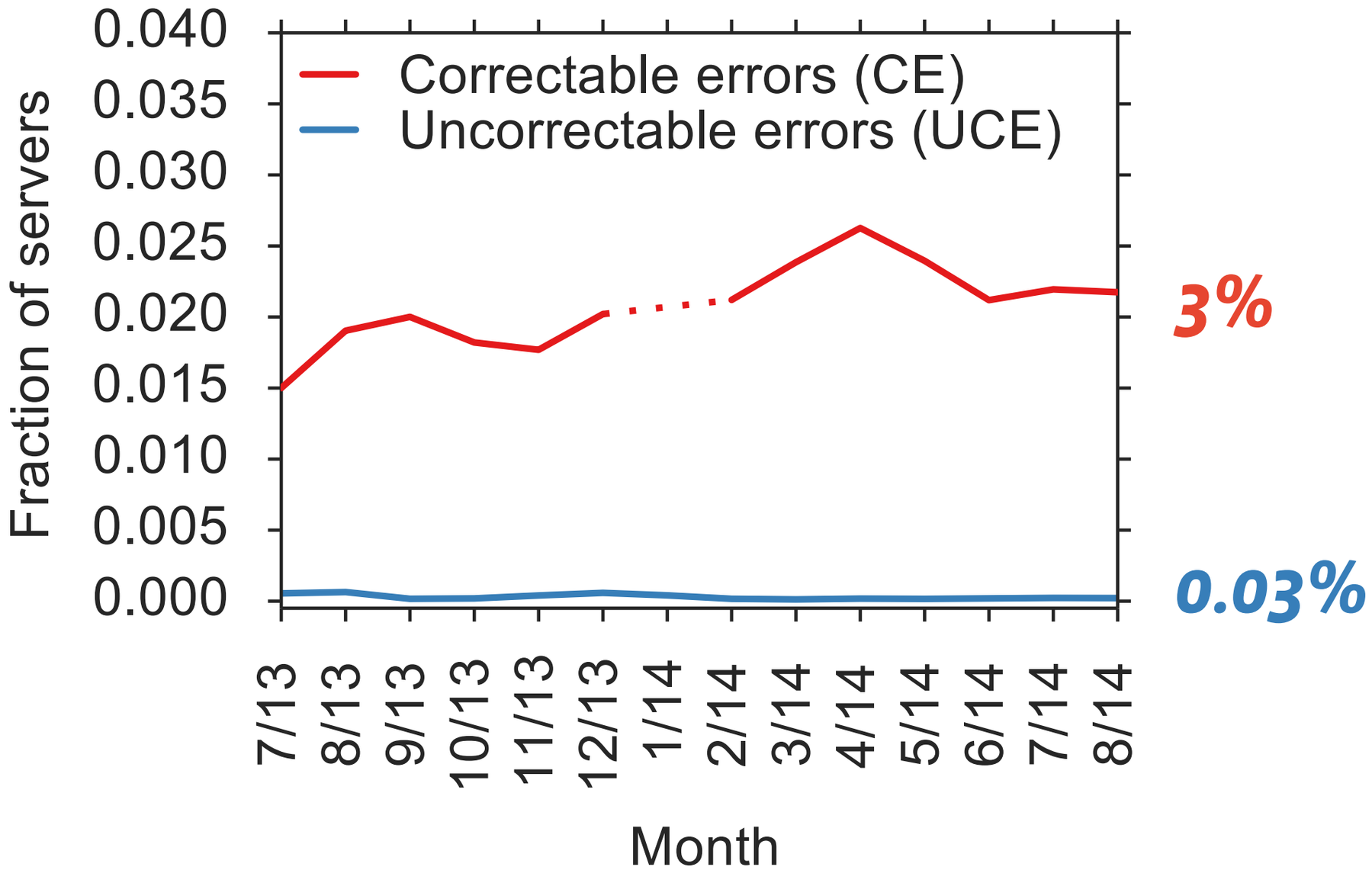
Technology scaling

Modeling errors

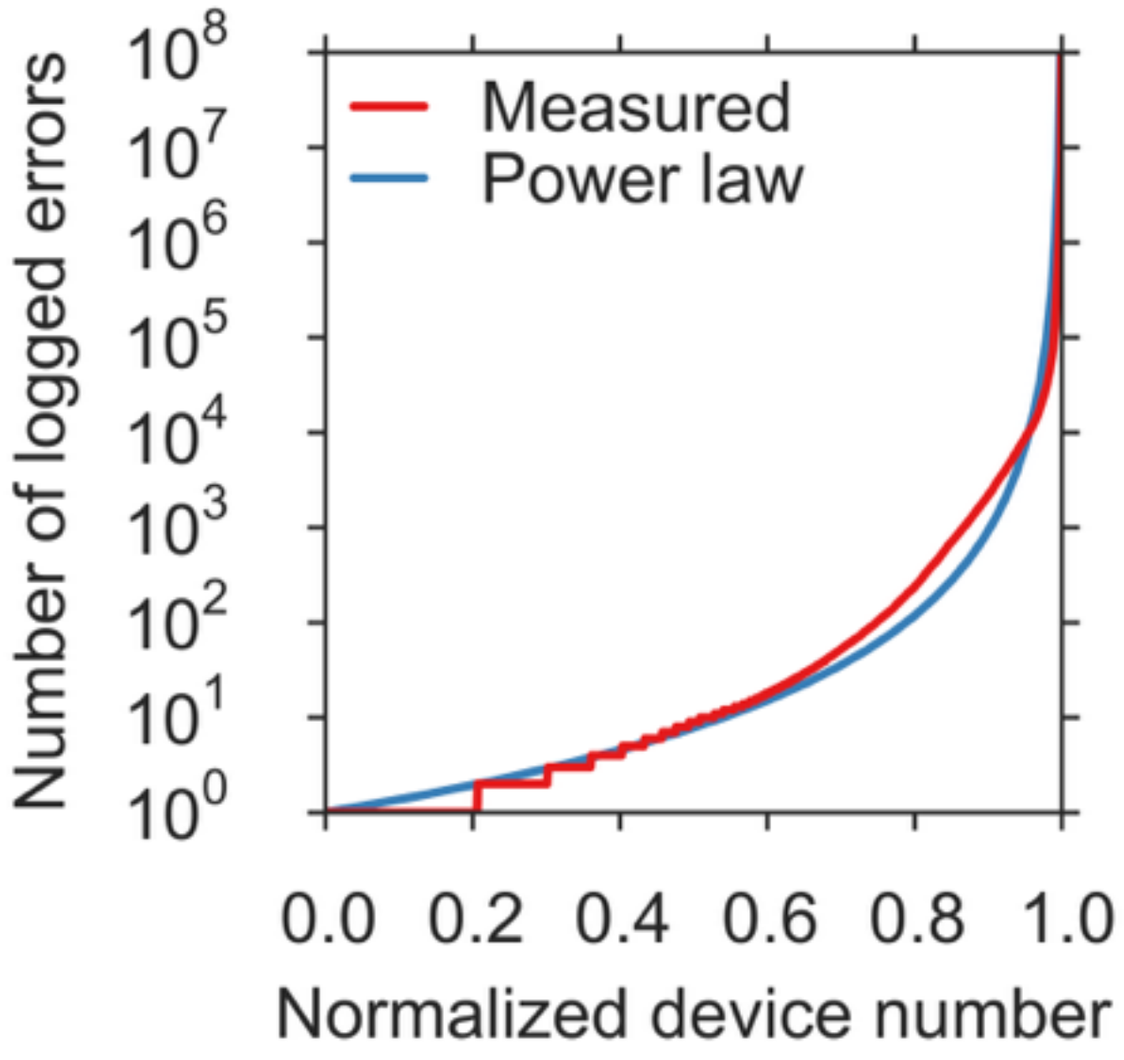
Architecture & workload

Error/failure occurrence

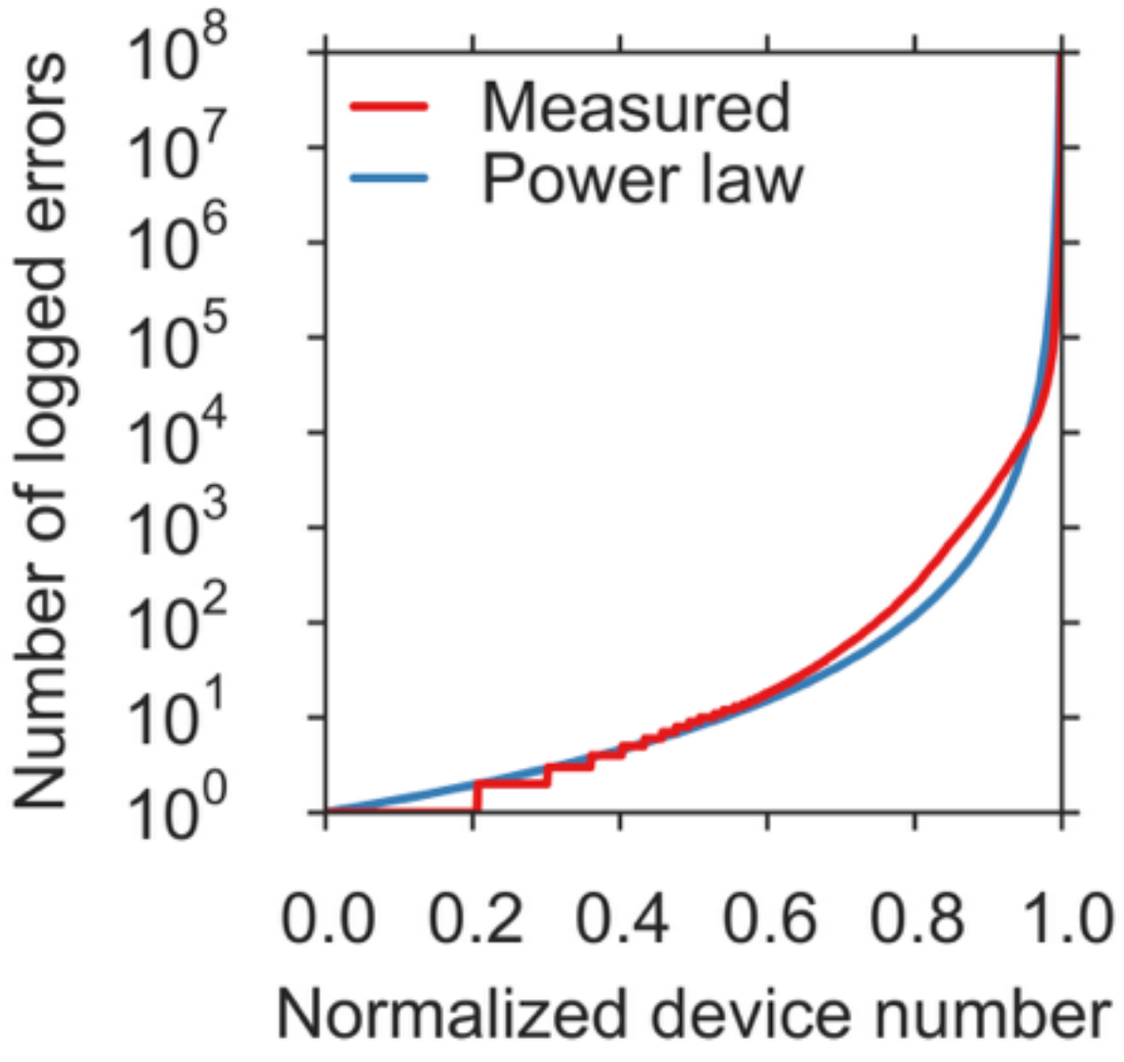
Server error rate



Memory error distribution

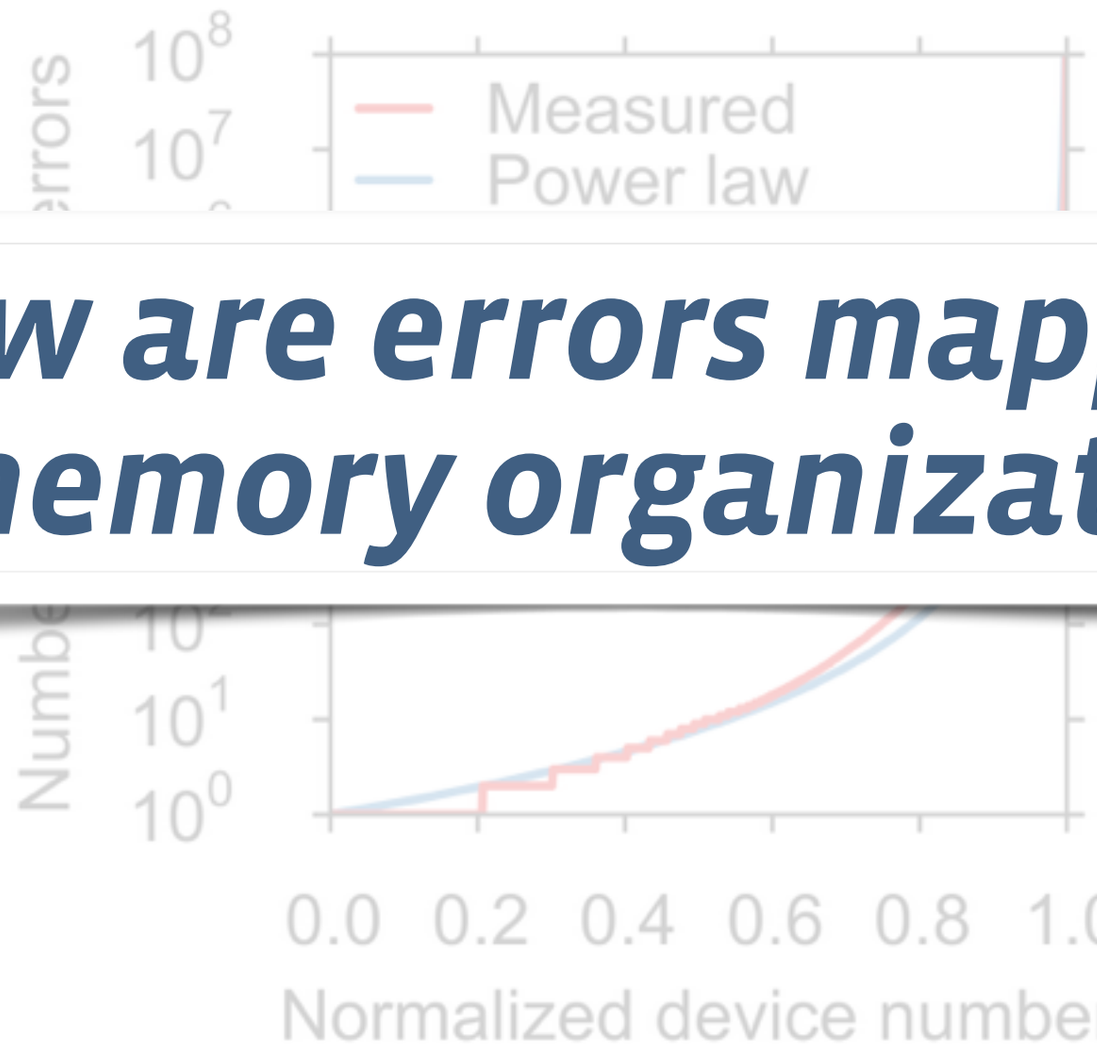


Memory error distribution



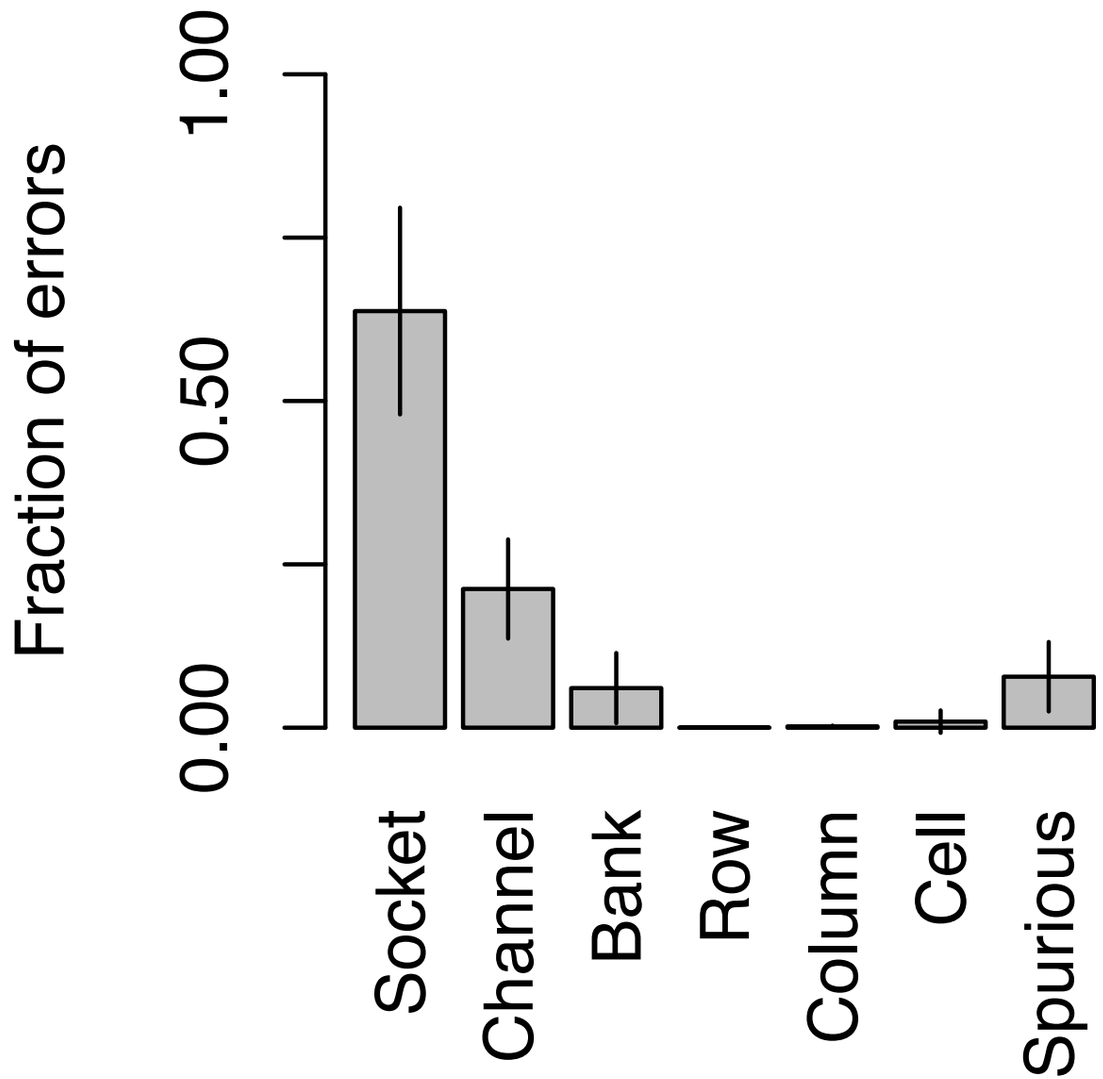
Decreasing hazard rate

Memory error distribution

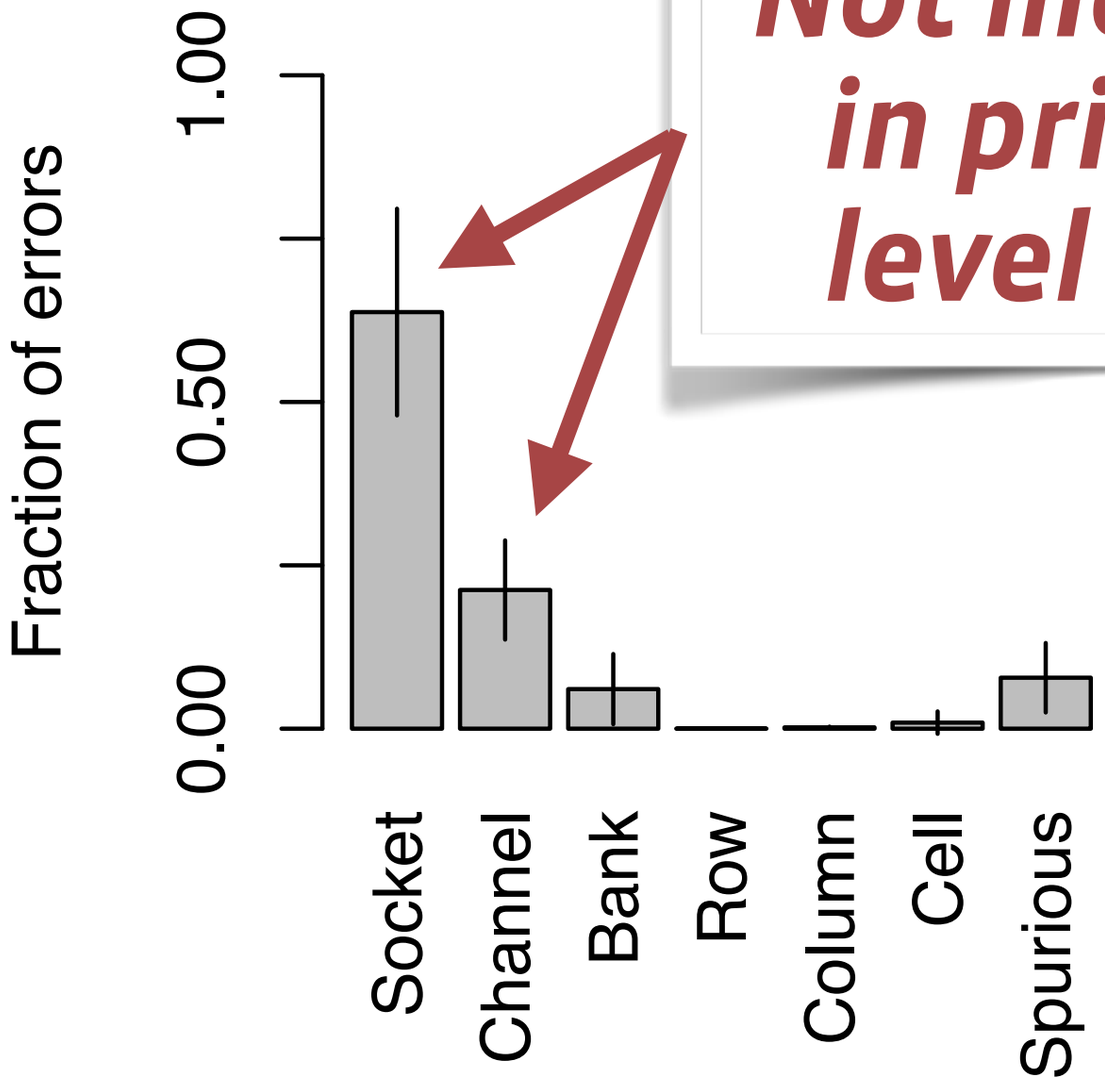


How are errors mapped to memory organization?

Sockets/channels: many errors

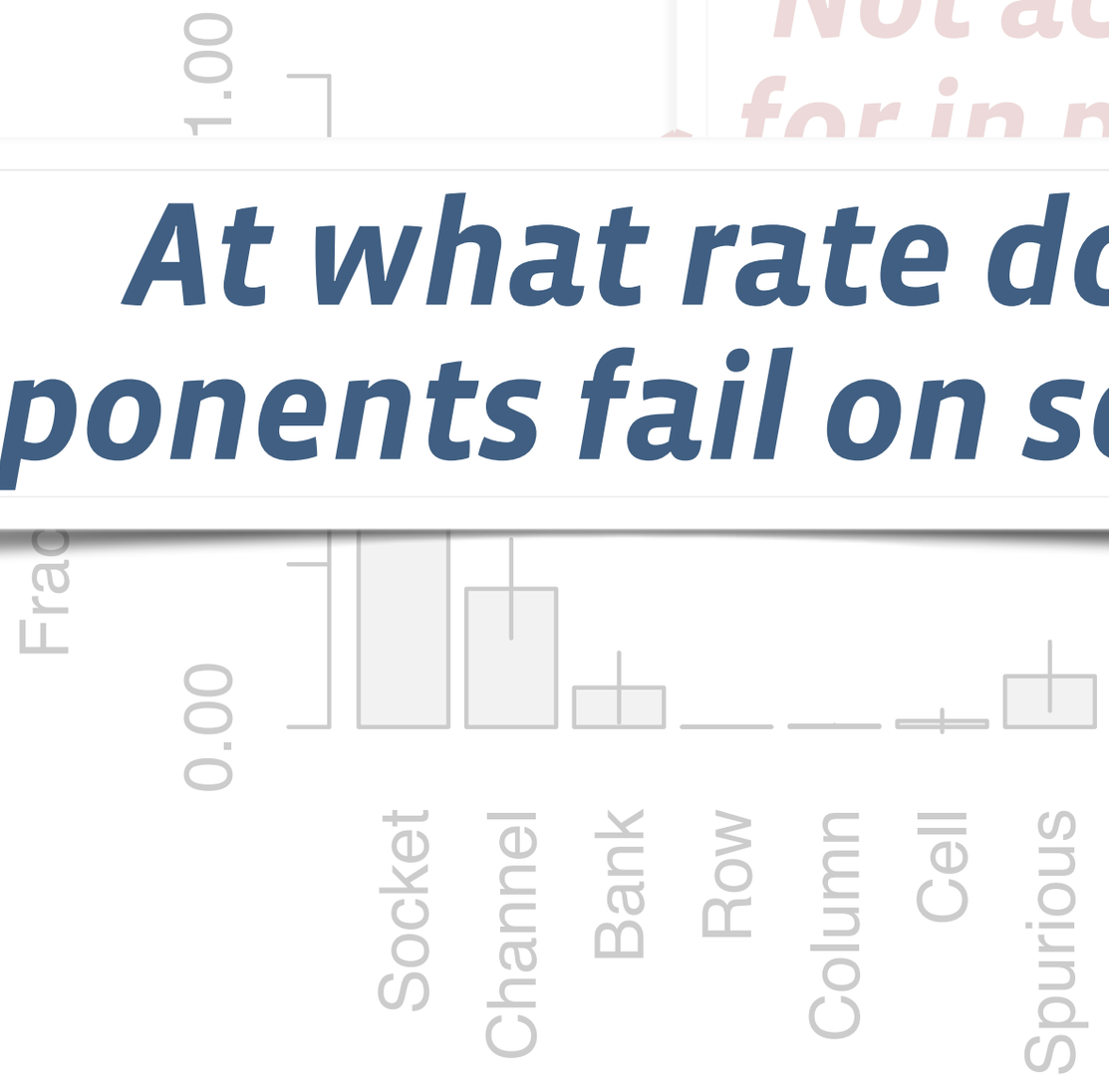


Sockets/channels: many errors



Not mentioned in prior chip-level studies

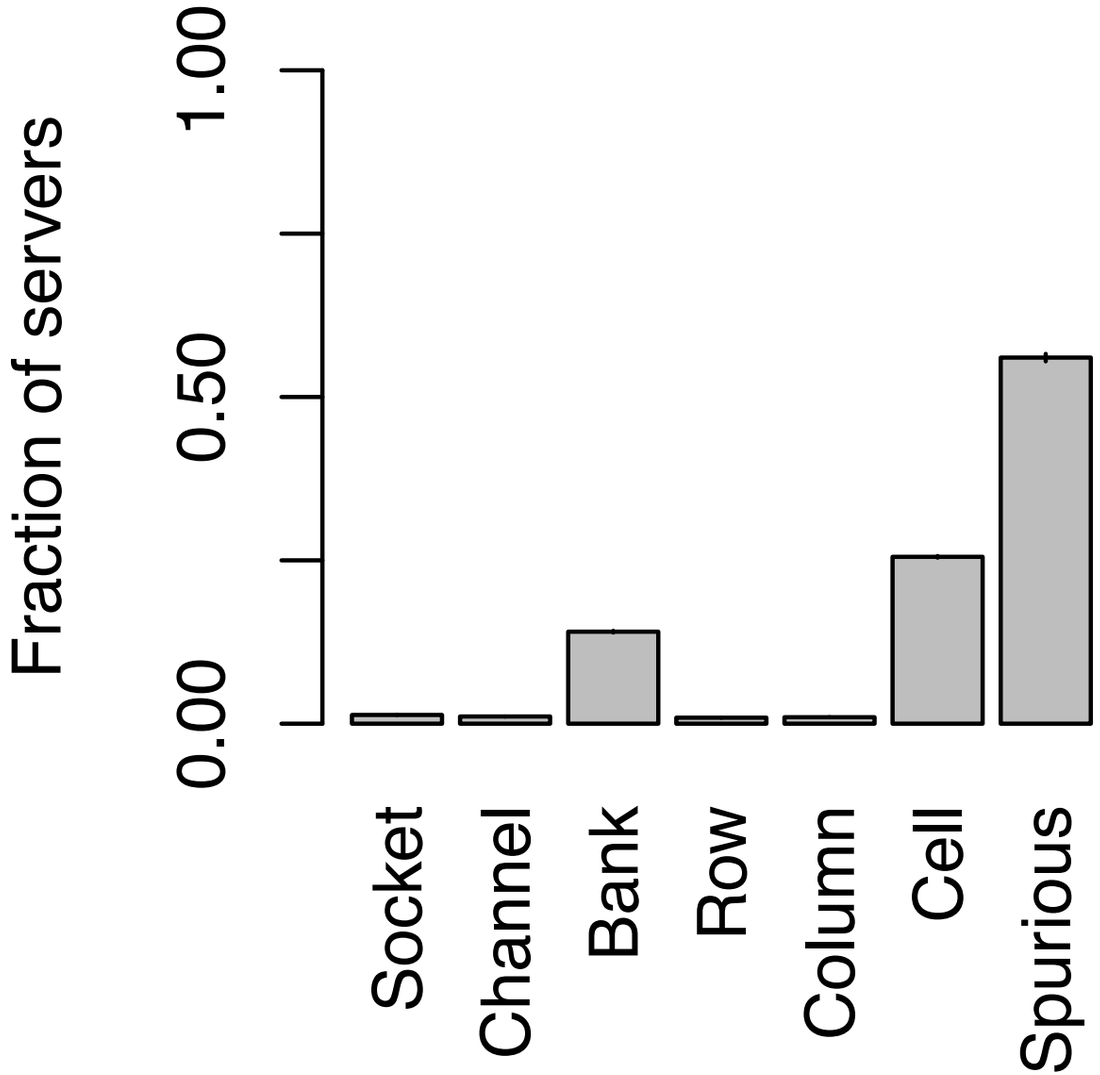
Sockets/channels: many errors



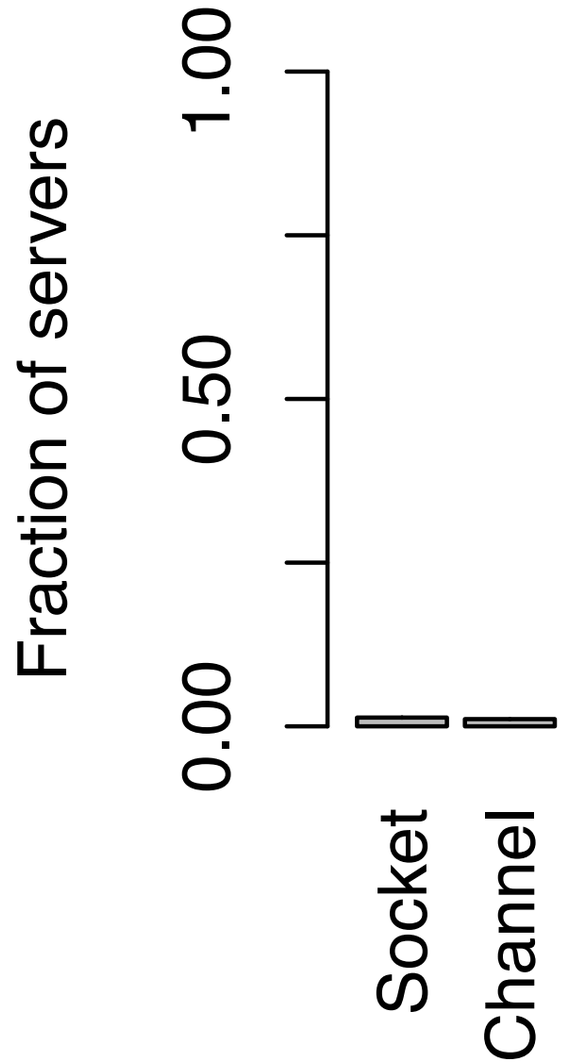
Not accounted for in prior chip-

At what rate do components fail on servers?

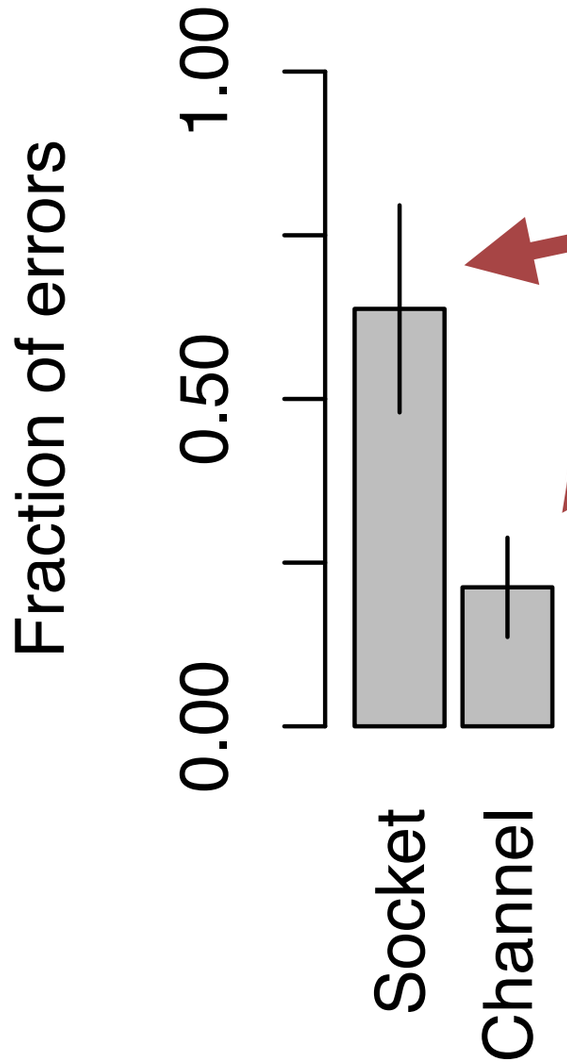
Bank/cell/spurious failures are common



of servers

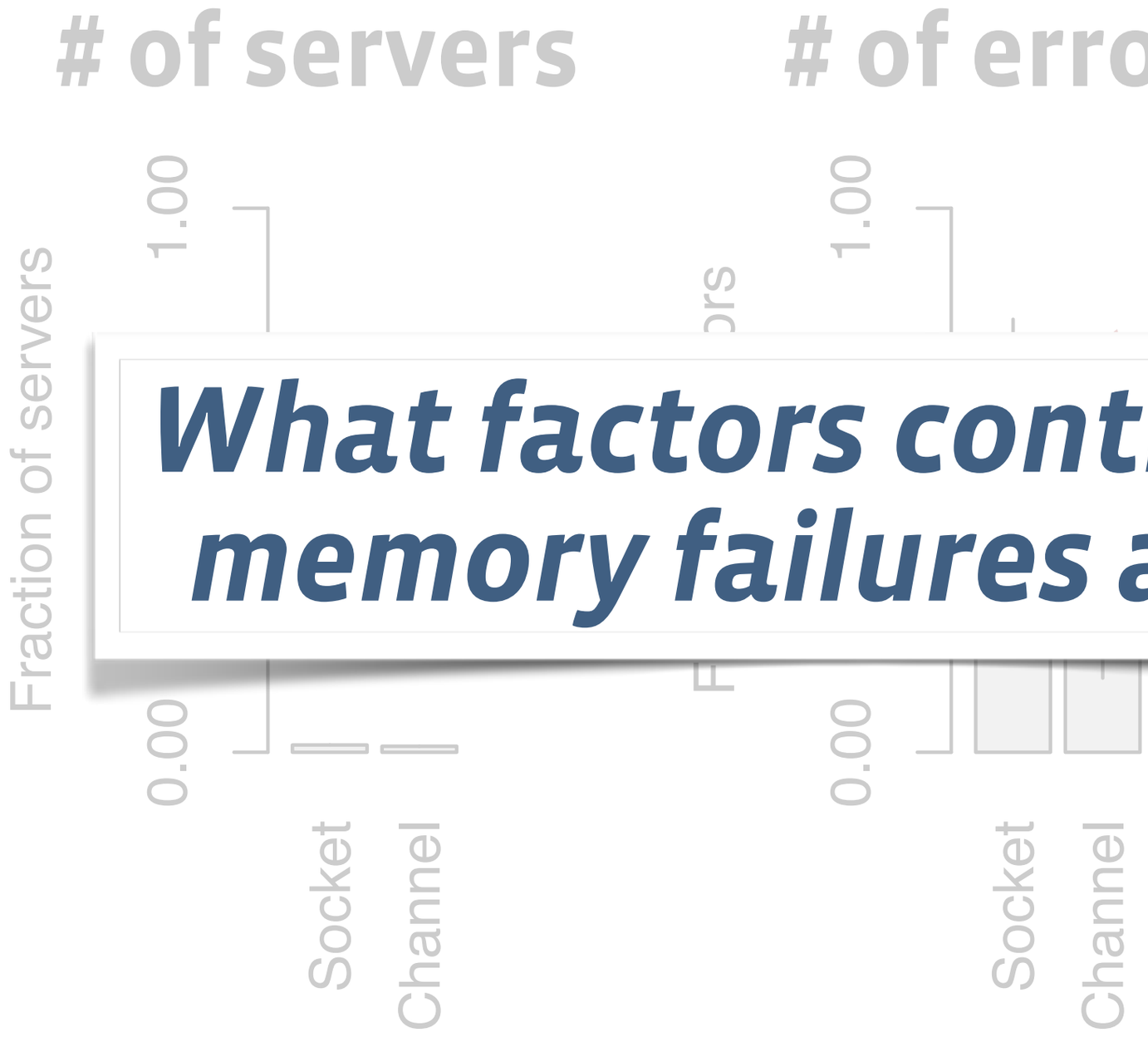


of errors



Denial-of-service–like behavior

What factors contribute to memory failures at scale?



Analytical methodology

- measure server characteristics
  - not feasible to examine every server
  - examined all servers ***with*** errors (error group)
  - ***sampled*** servers ***without*** errors (control group)
- bucket devices based on characteristics
- measure ***relative failure rate***
  - of error group vs. control group
  - within each bucket

Error/failure occurrence

Page offlining at scale

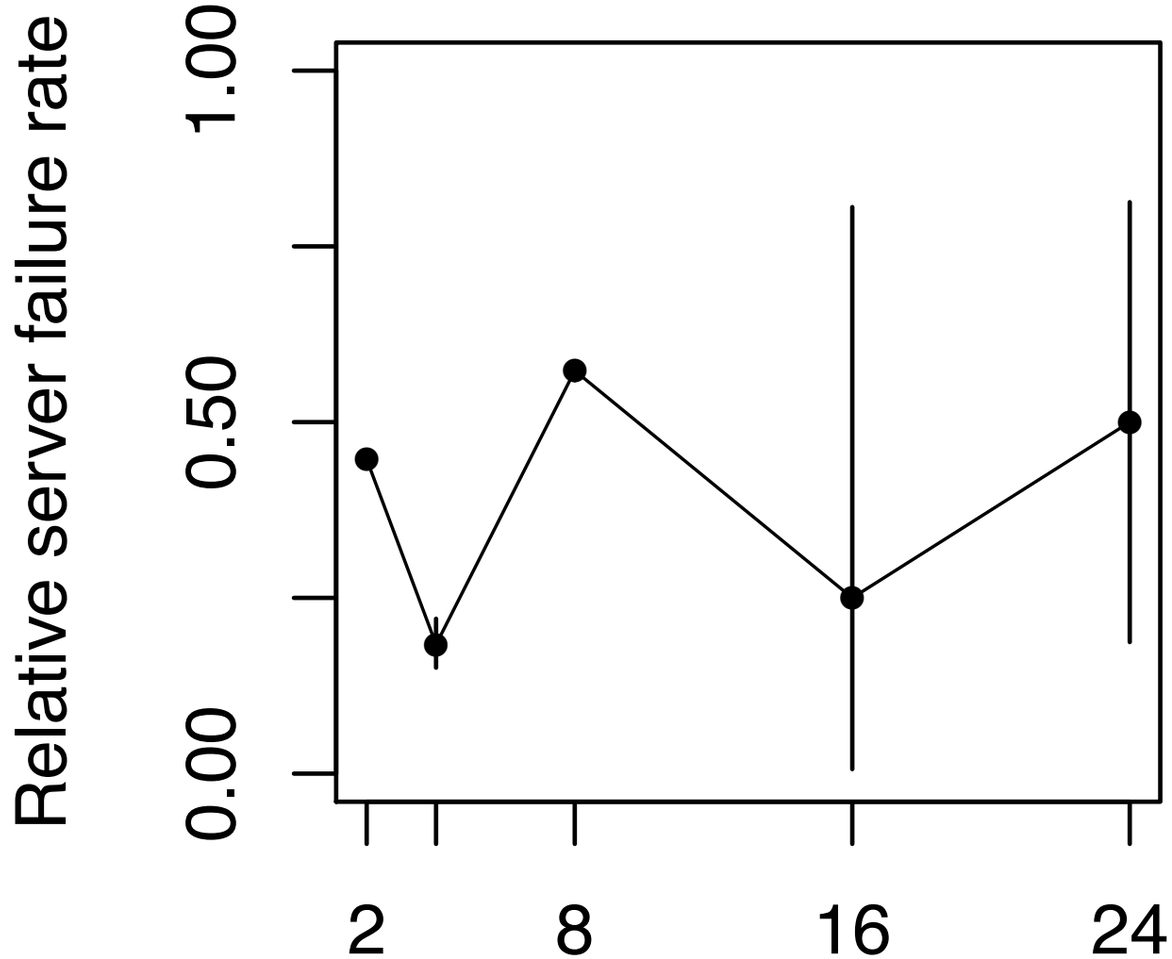
New reliability trends

Technology scaling

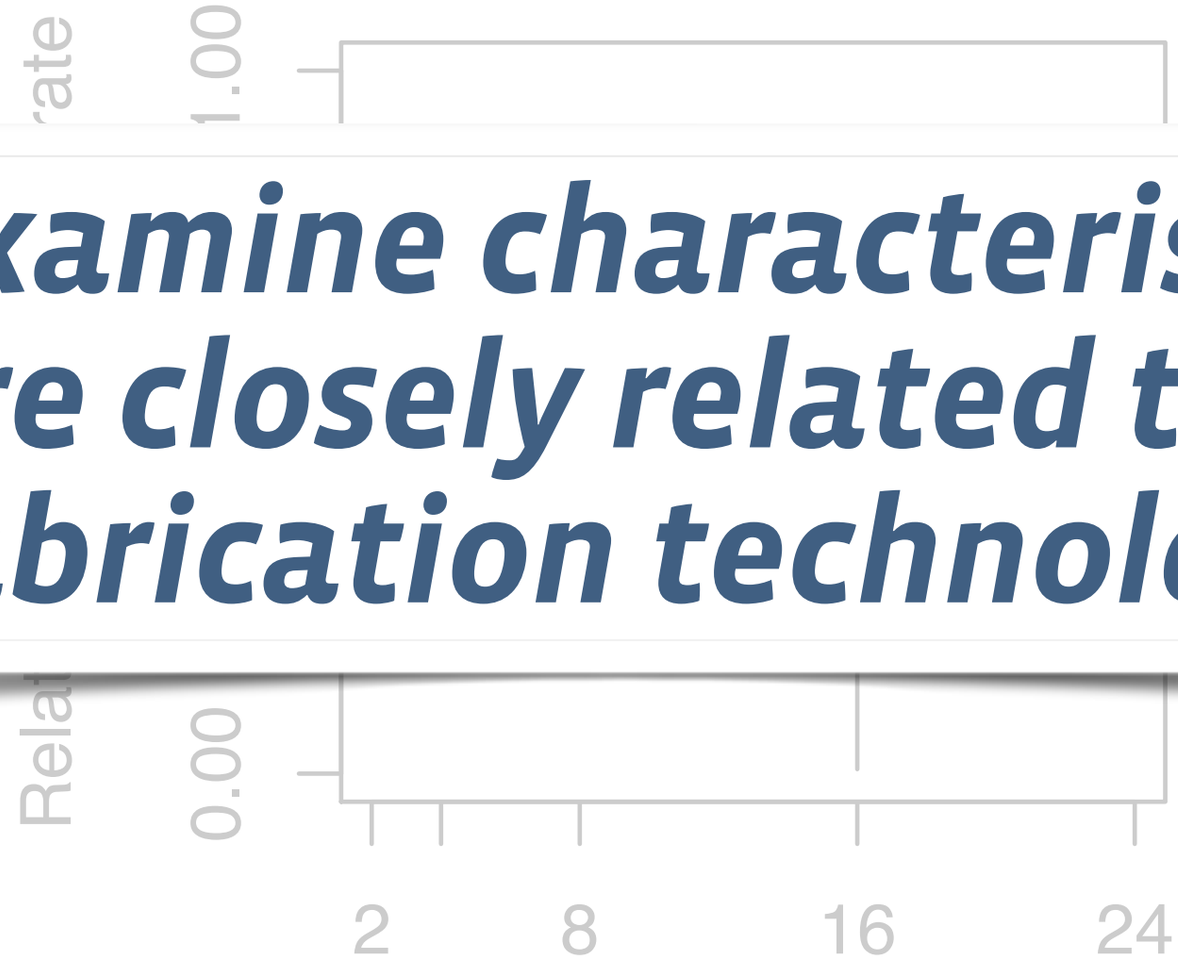
Modeling errors

Architecture & workload

Prior work found *inconclusive trends* with respect to memory ***capacity***



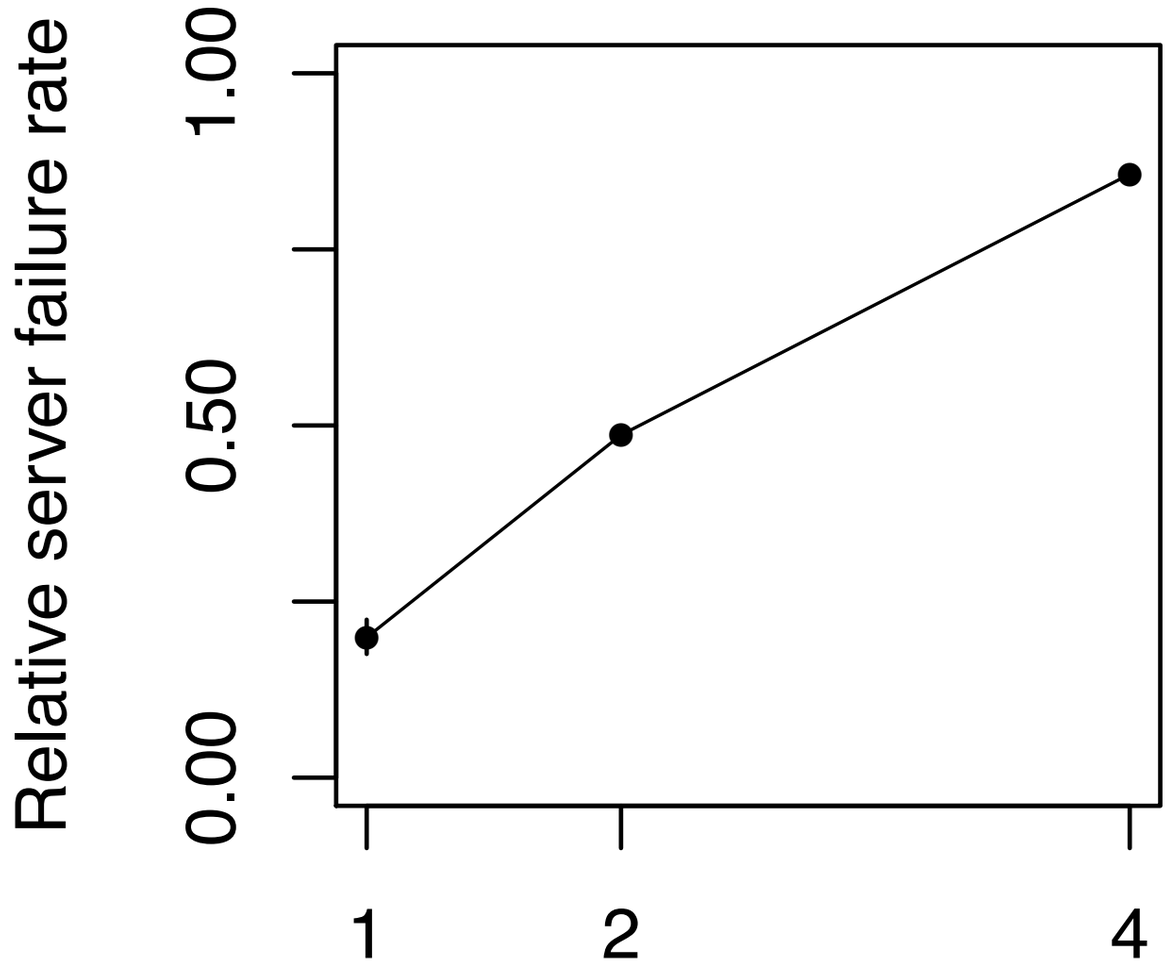
Prior work found ***inconclusive trends*** with respect to memory ***capacity***

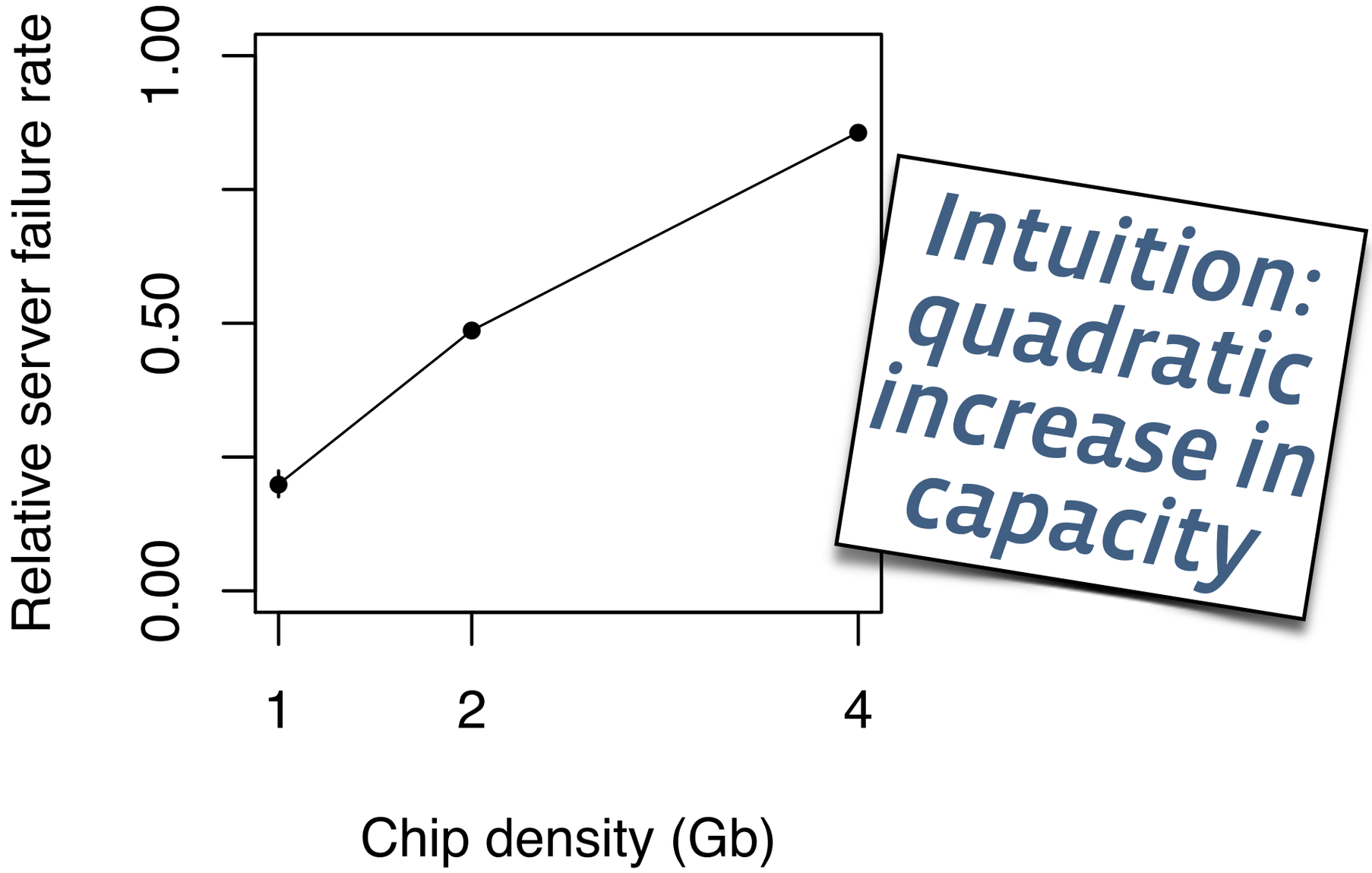


Examine characteristic more closely related to cell fabrication technology

Use **DRAM chip density**
to examine technology scaling

(closely related to fabrication technology)


Relative server failure rate
0.00
0.50
1.00
1
2
4
Chip density (Gb)


Relative server failure rate
0.00
0.50
1.00
1
2
4
Chip density (Gb)
Intuition: quadratic increase in capacity

Error/failure occurrence

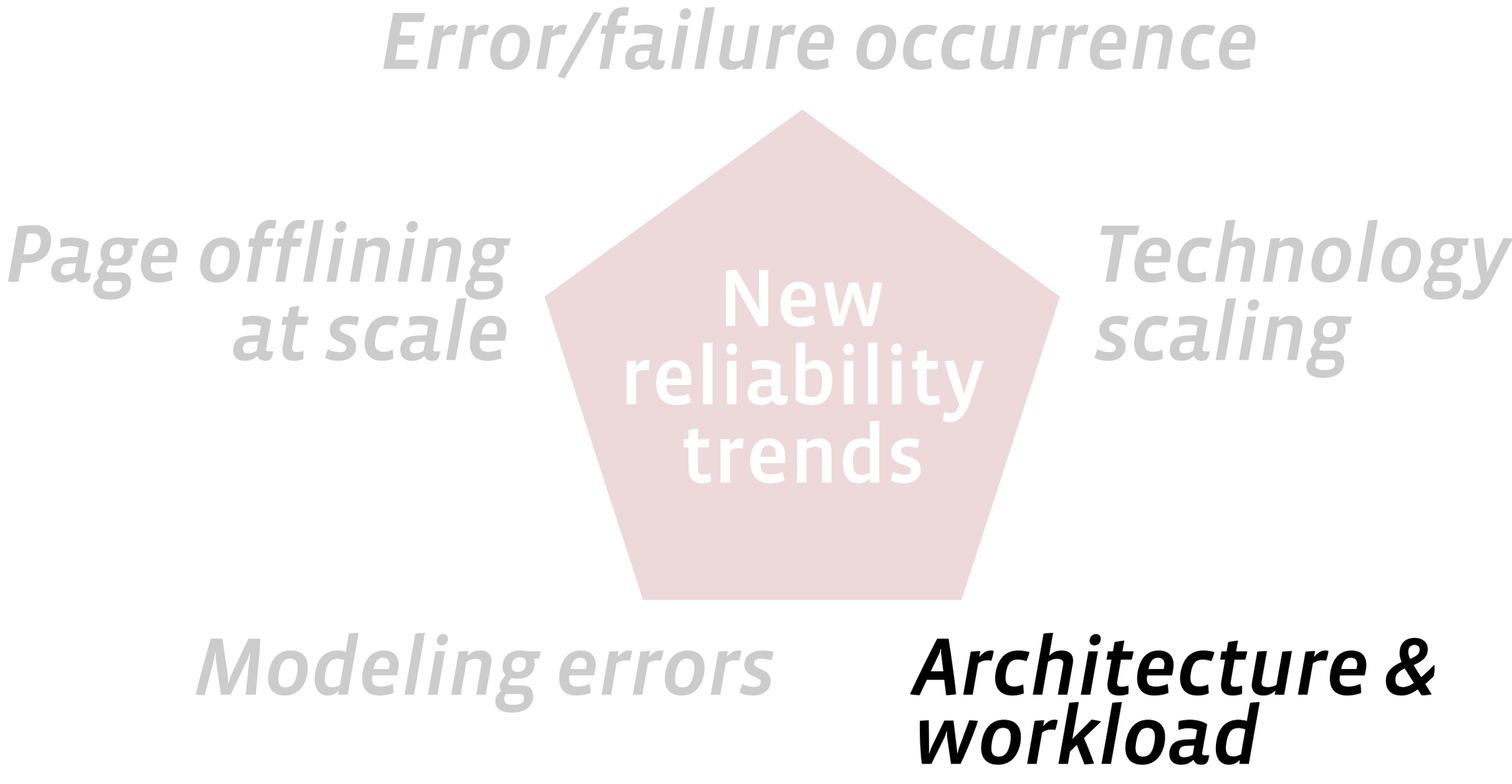
We find that ***newer*** cell fabrication technologies have ***higher failure rates***

Technology scaling

trends

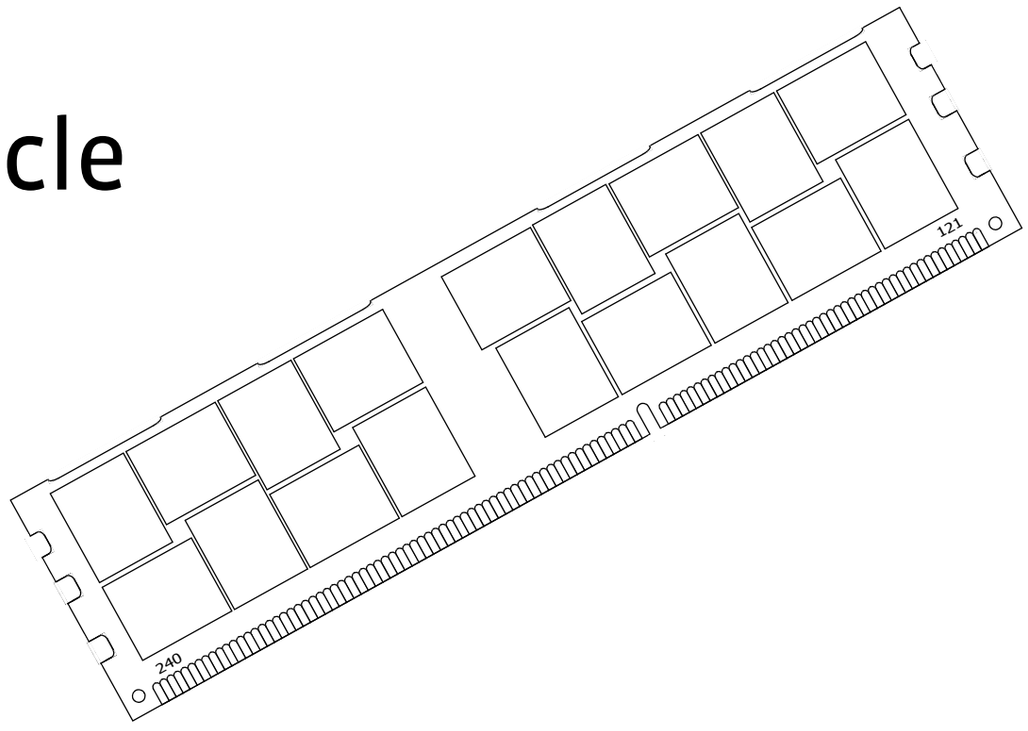
Modeling errors

Architecture & workload


Error/failure occurrence
Page offlining at scale
New reliability trends
Technology scaling
Modeling errors
Architecture & workload

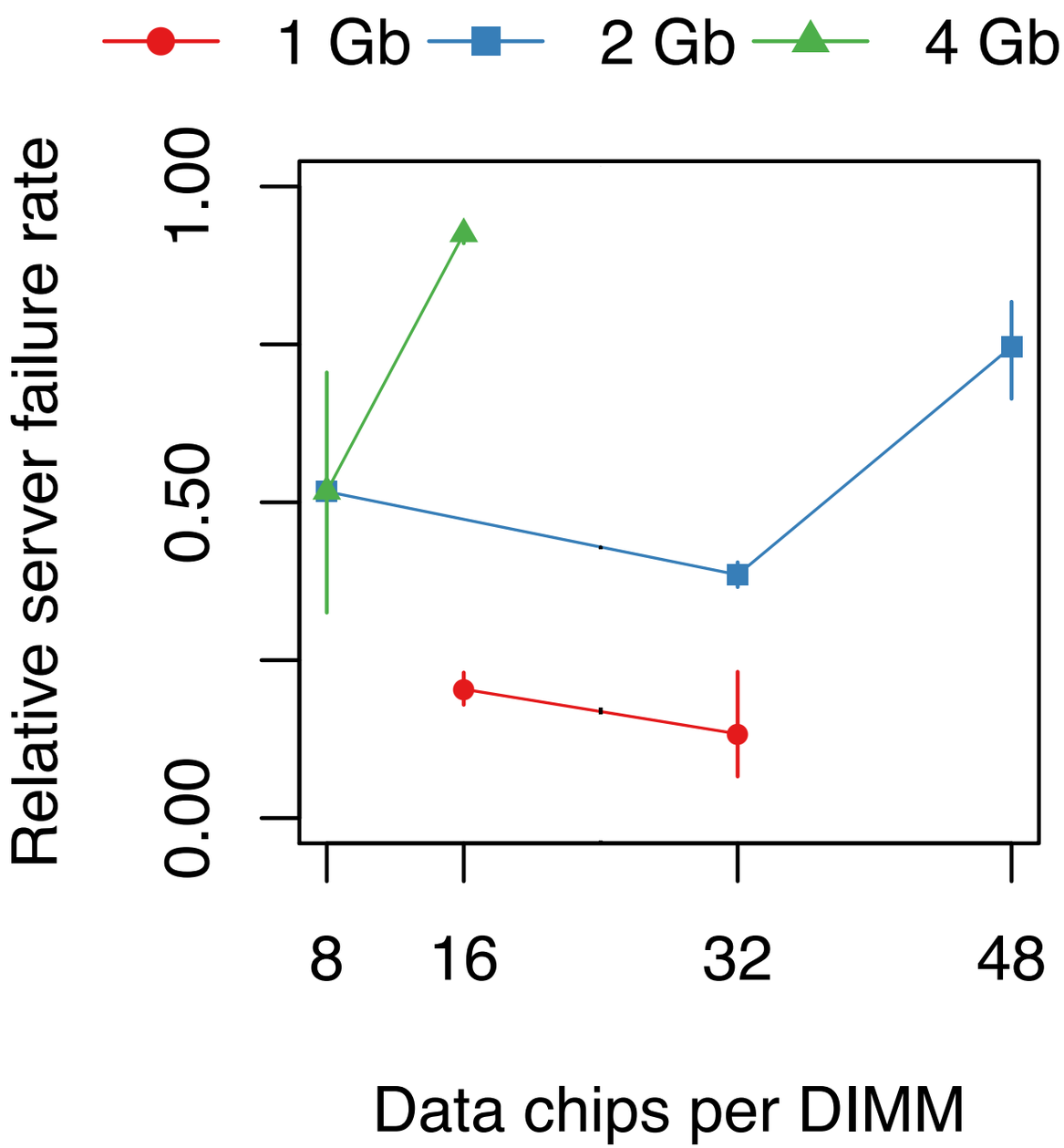
DIMM architecture

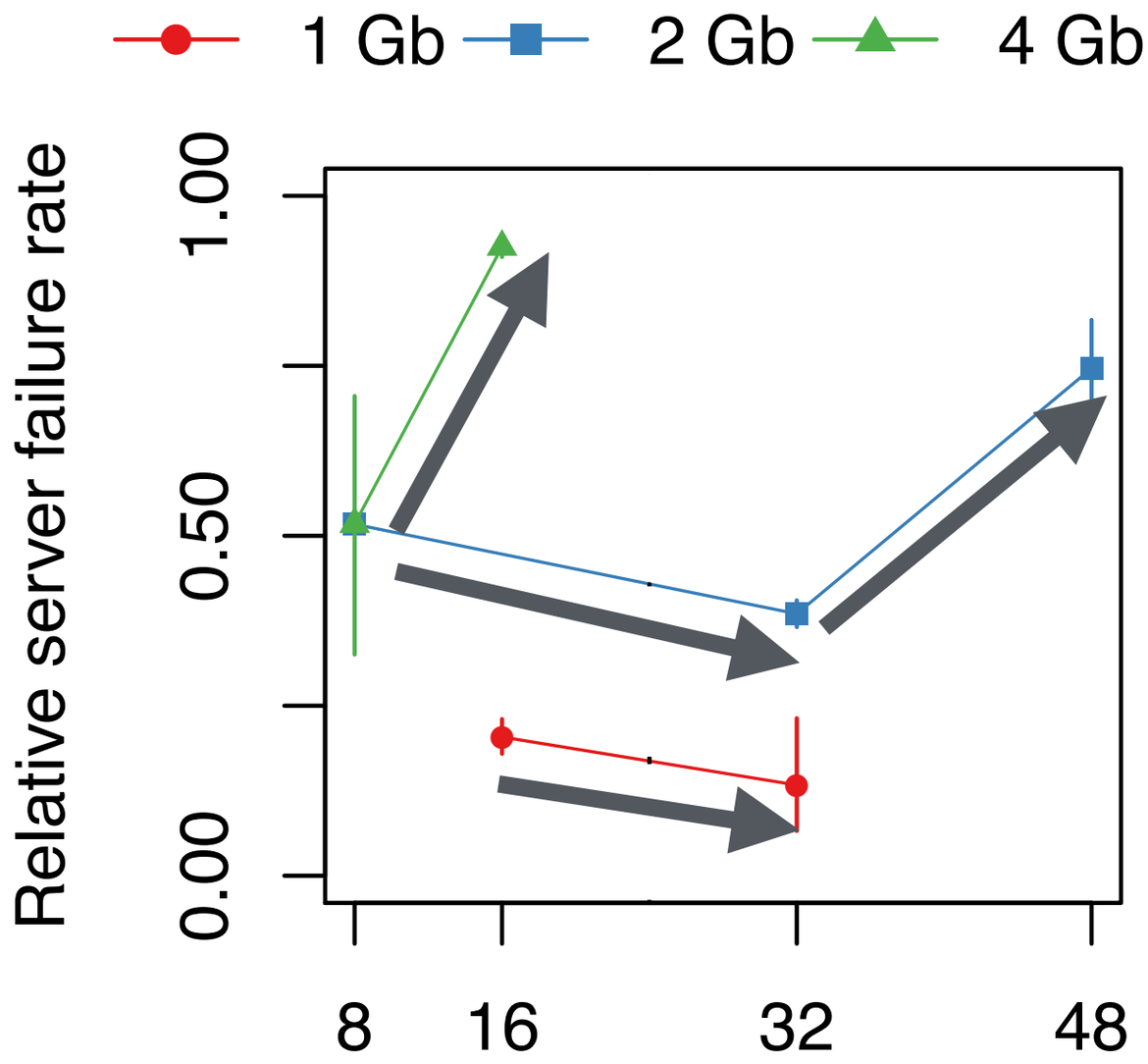
- chips per DIMM, transfer width
  - 8 to 48 chips
  - x4, x8 = 4 or 8 bits per cycle
- electrical implications

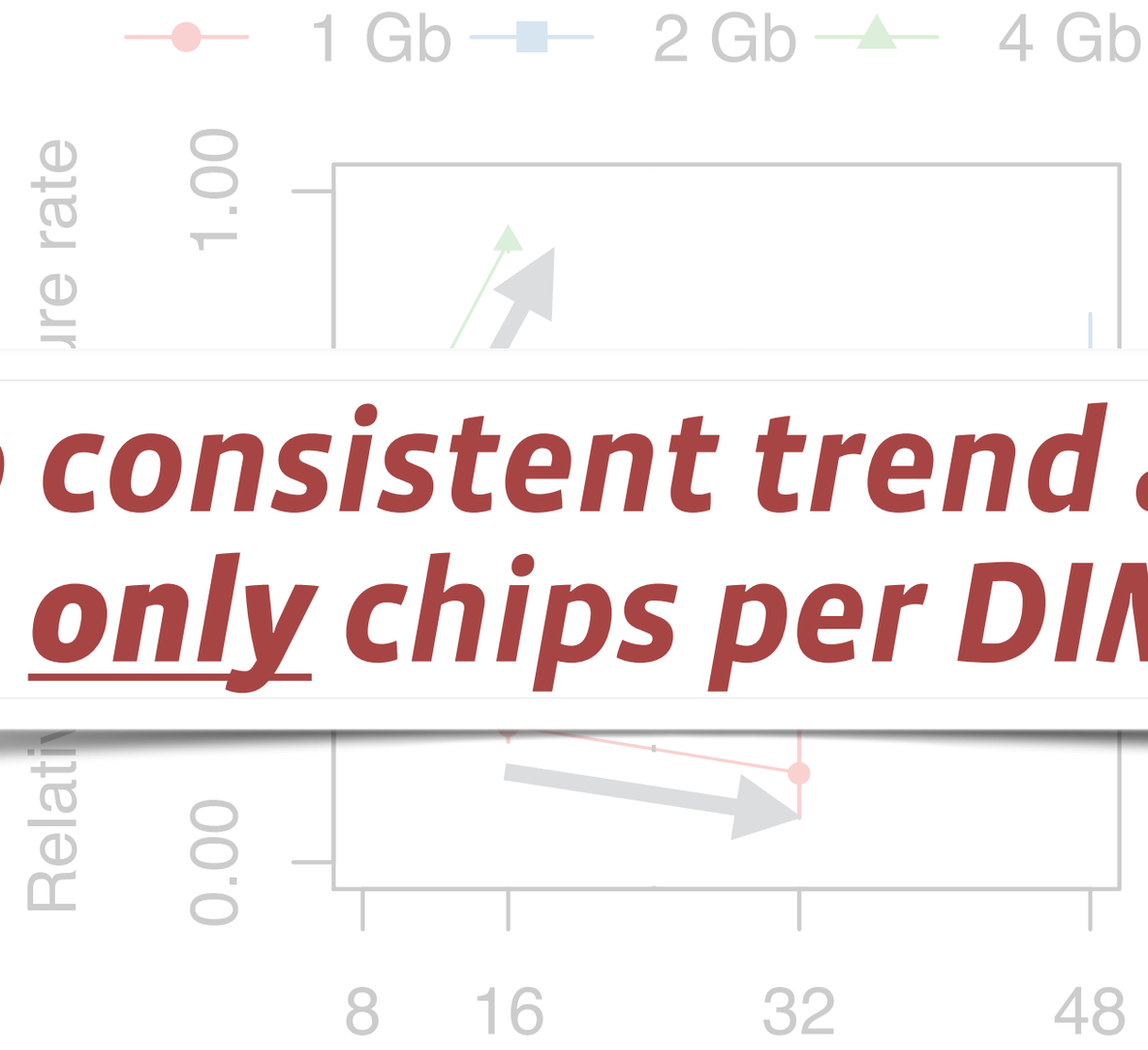


DIMM architecture

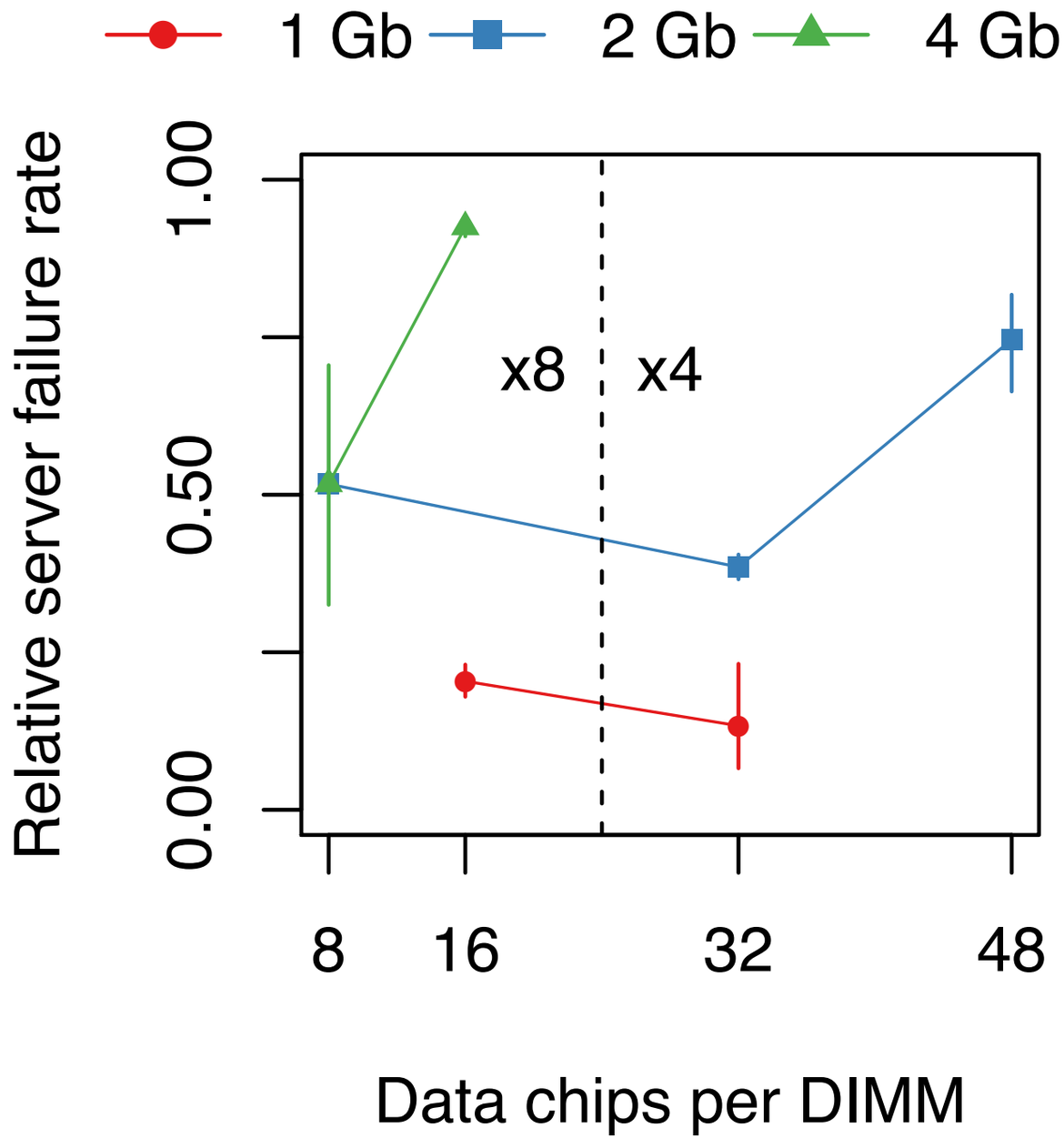
- ***Does DIMM organization affect memory reliability?***
- electrical implications

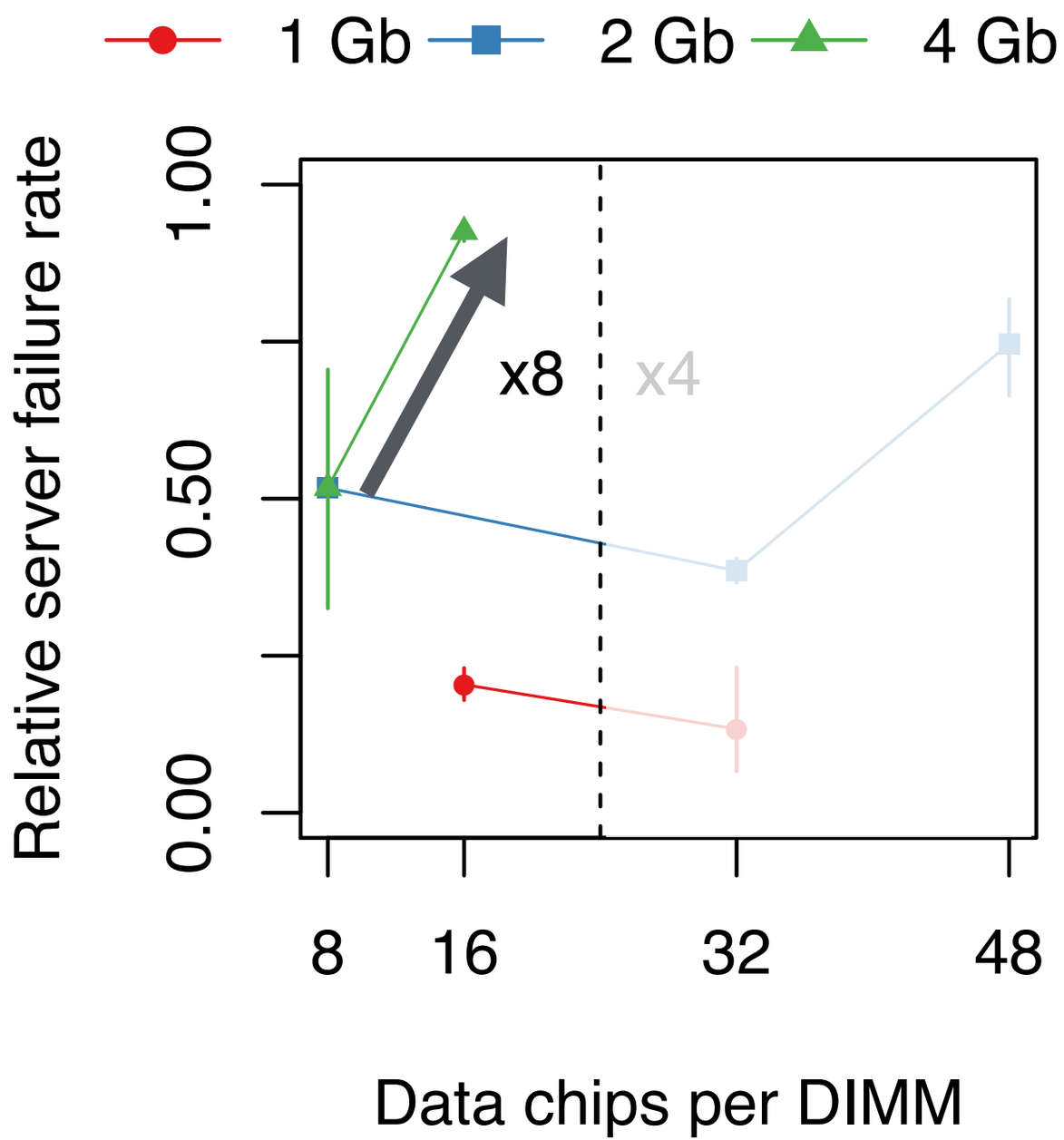

1 Gb
2 Gb
4 Gb
Relative server failure rate
1.00
0.50
0.00
8
16
32
48
Data chips per DIMM

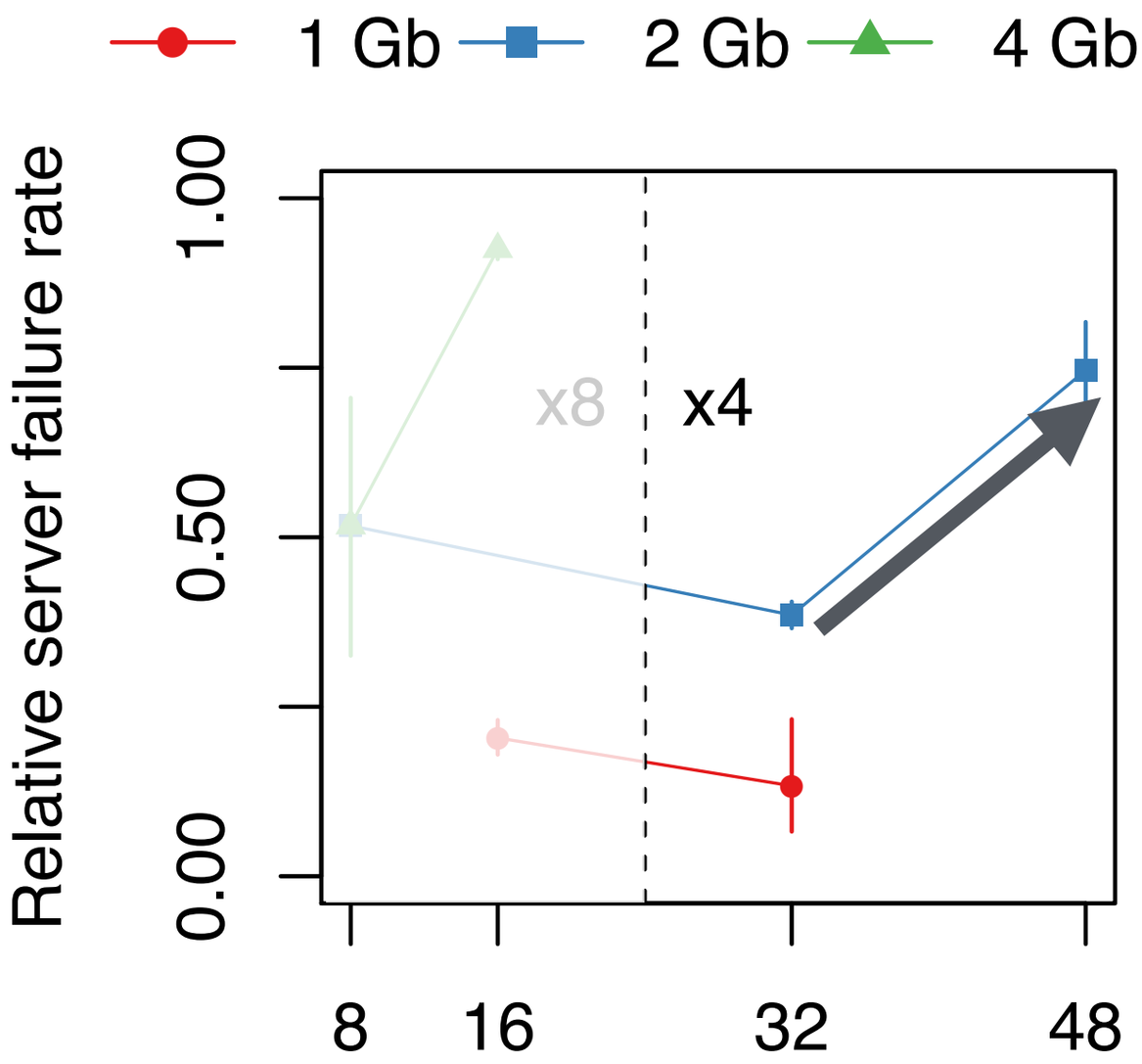

1 Gb
2 Gb
4 Gb
Relative server failure rate
0.00
0.50
1.00
8
16
32
48
Data chips per DIMM

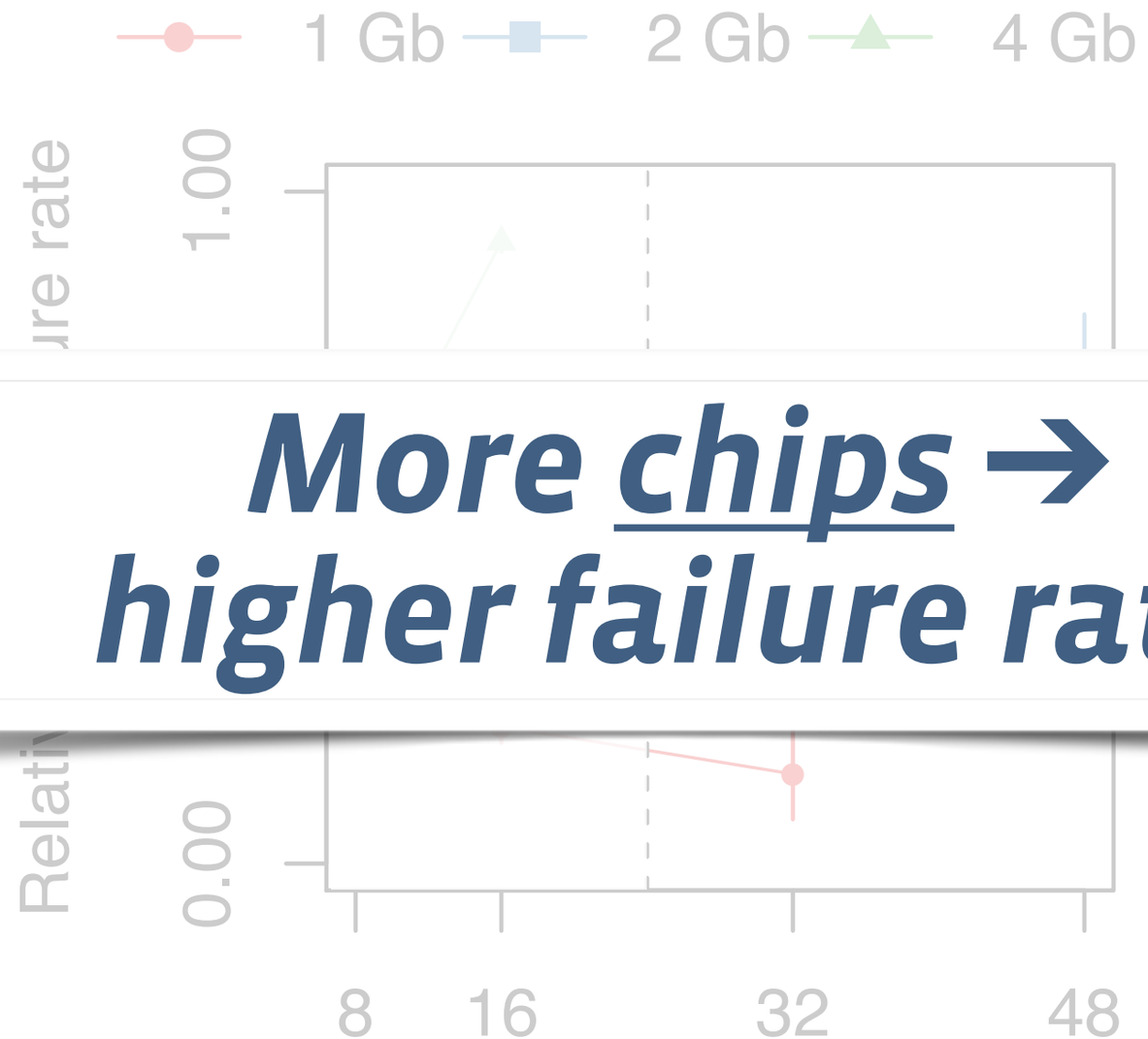


No consistent trend across <u>only</u> chips per DIMM


1 Gb
2 Gb
4 Gb
Relative server failure rate
1.00
0.50
0.00
x8
x4
8
16
32
48
Data chips per DIMM

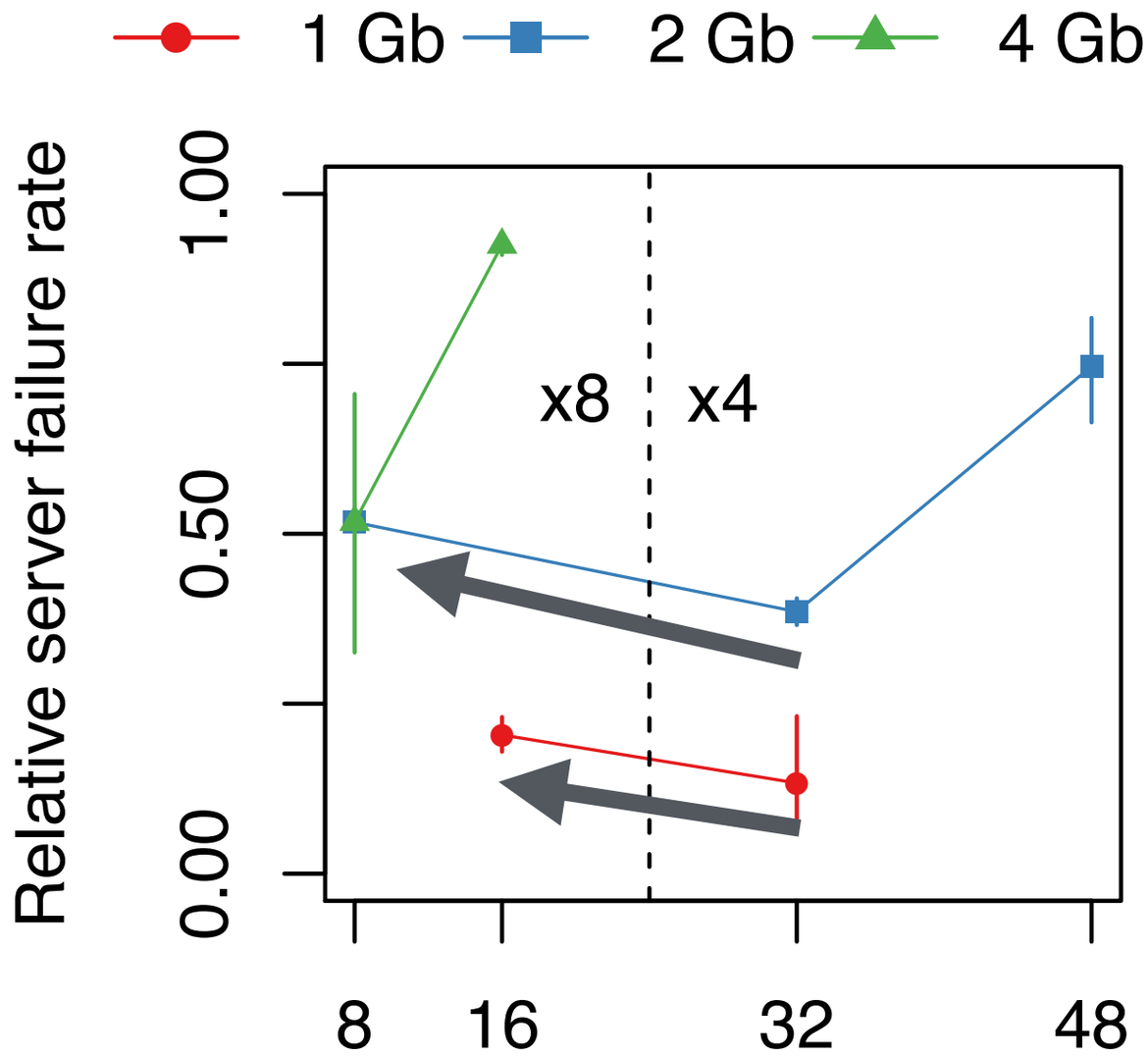

1 Gb
2 Gb
4 Gb
Relative server failure rate
1.00
0.50
0.00
x8
x4
8
16
32
48
Data chips per DIMM

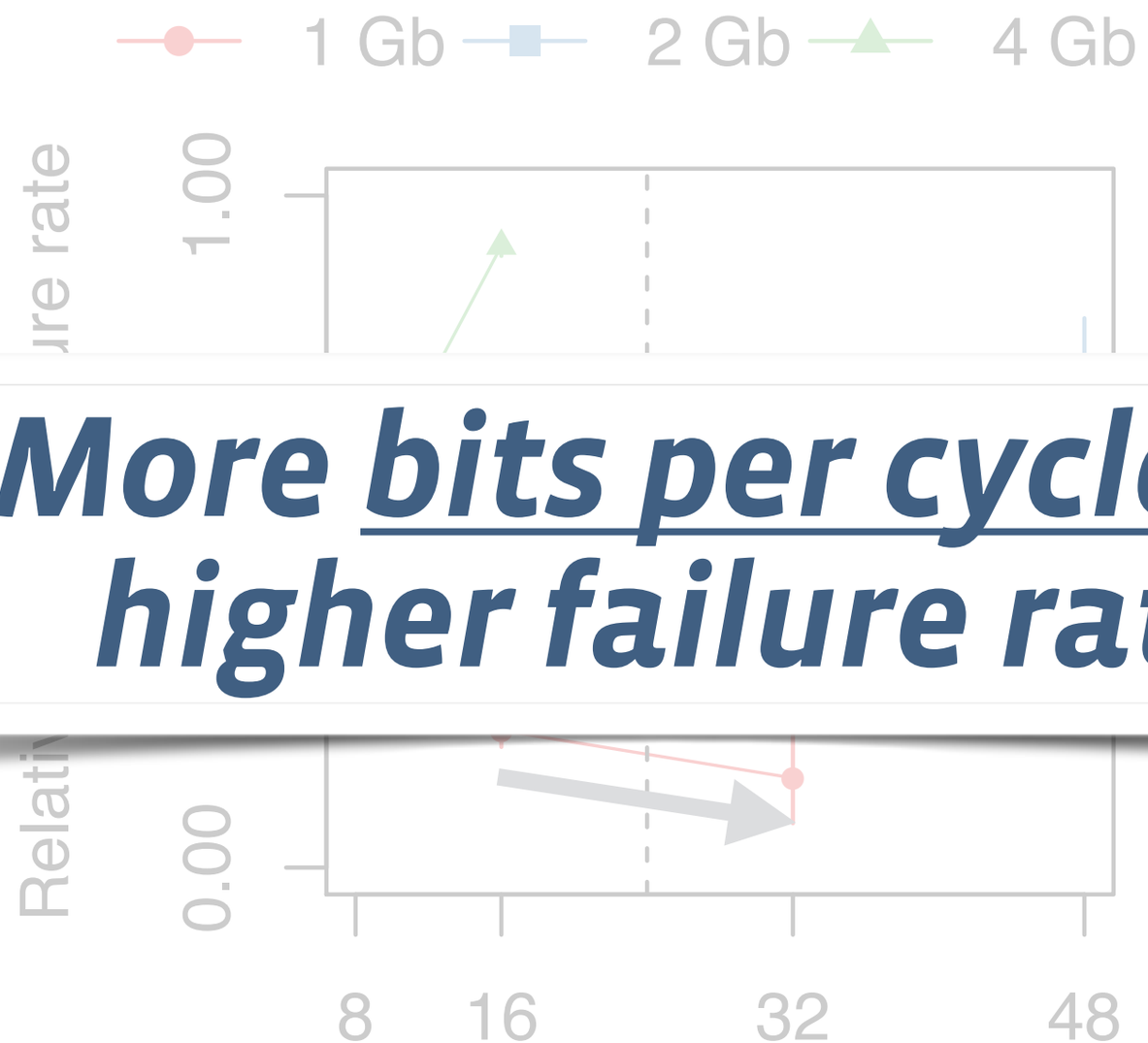

1 Gb
2 Gb
4 Gb
Relative server failure rate
1.00
0.50
0.00
x8
x4
8
16
32
48
Data chips per DIMM



More chips →

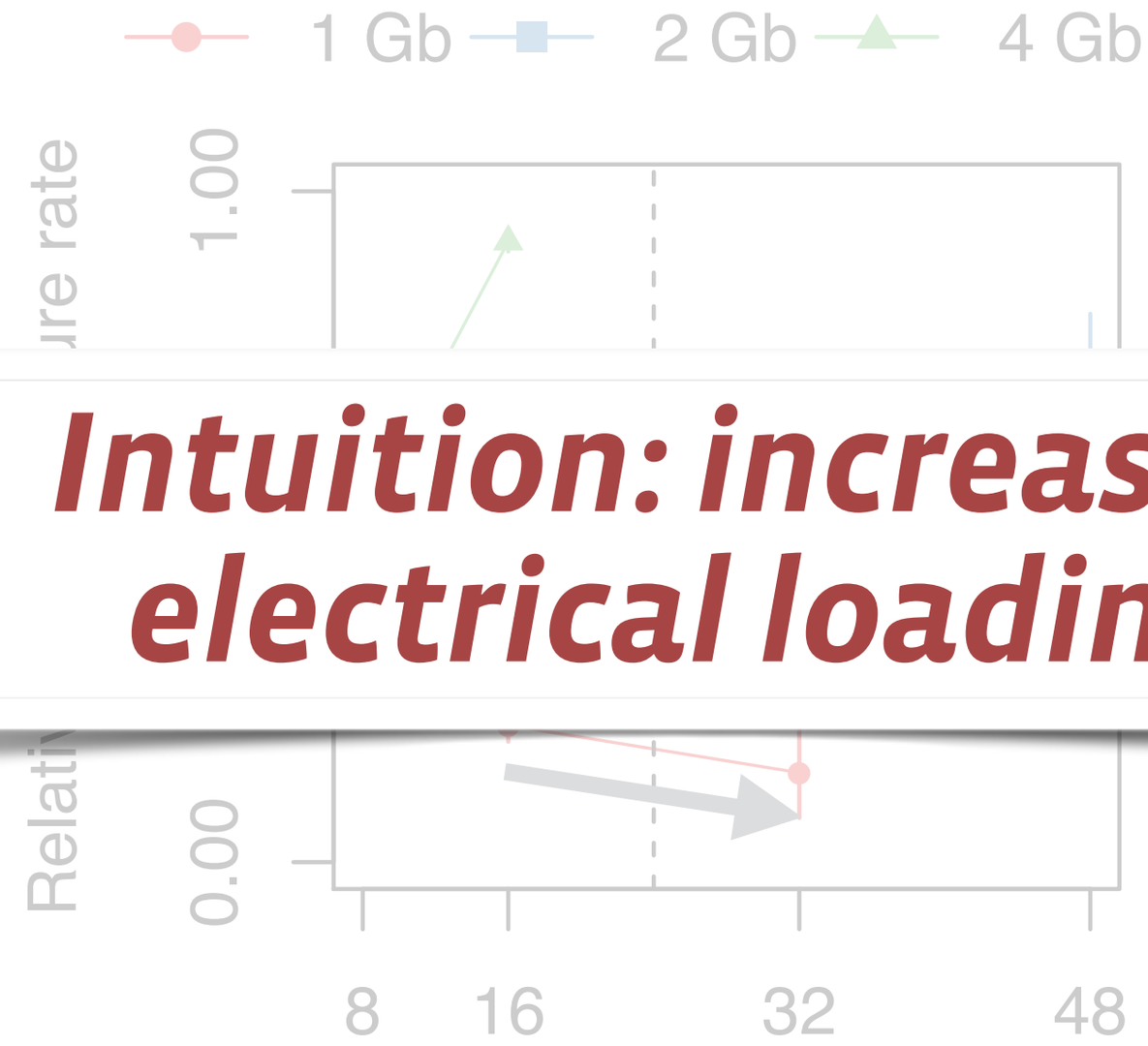
higher failure rate


1 Gb
2 Gb
4 Gb
Relative server failure rate
1.00
0.50
0.00
x8
x4
8
16
32
48
Data chips per DIMM



More bits per cycle →

higher failure rate



Intuition: increased electrical loading

Workload dependence

- prior studies: homogeneous workloads
  - web search and scientific
- warehouse-scale data centers:
  - web, hadoop, ingest, database, cache, media

Workload dependence

- prior studies: homogeneous workloads

What affect to <u>heterogeneous</u> workloads have on reliability?

dia

1 Gb 2 Gb 4 Gb

Relative server failure rate

1.00
0.50
0.00

0 0.25 0.5 0.75 1

CPU utilization

1.00
0.50
0.00

0 0.25 0.5 0.75 1

Memory utilization

1 Gb 2 Gb 4 Gb

Relative server failure rate

1.00

0.00

0 0.25 0.5 0.75 1

CPU utilization

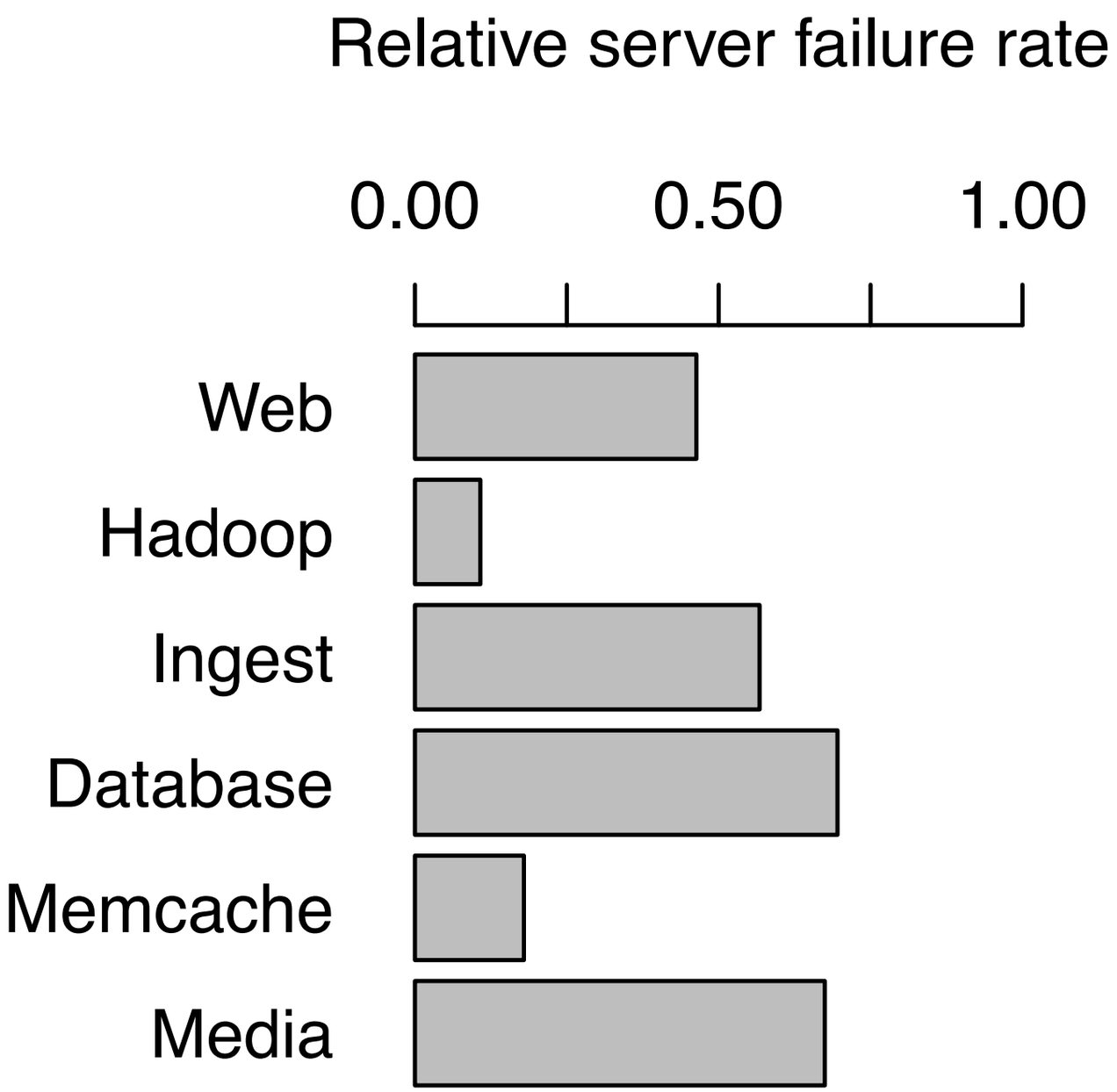
1.00

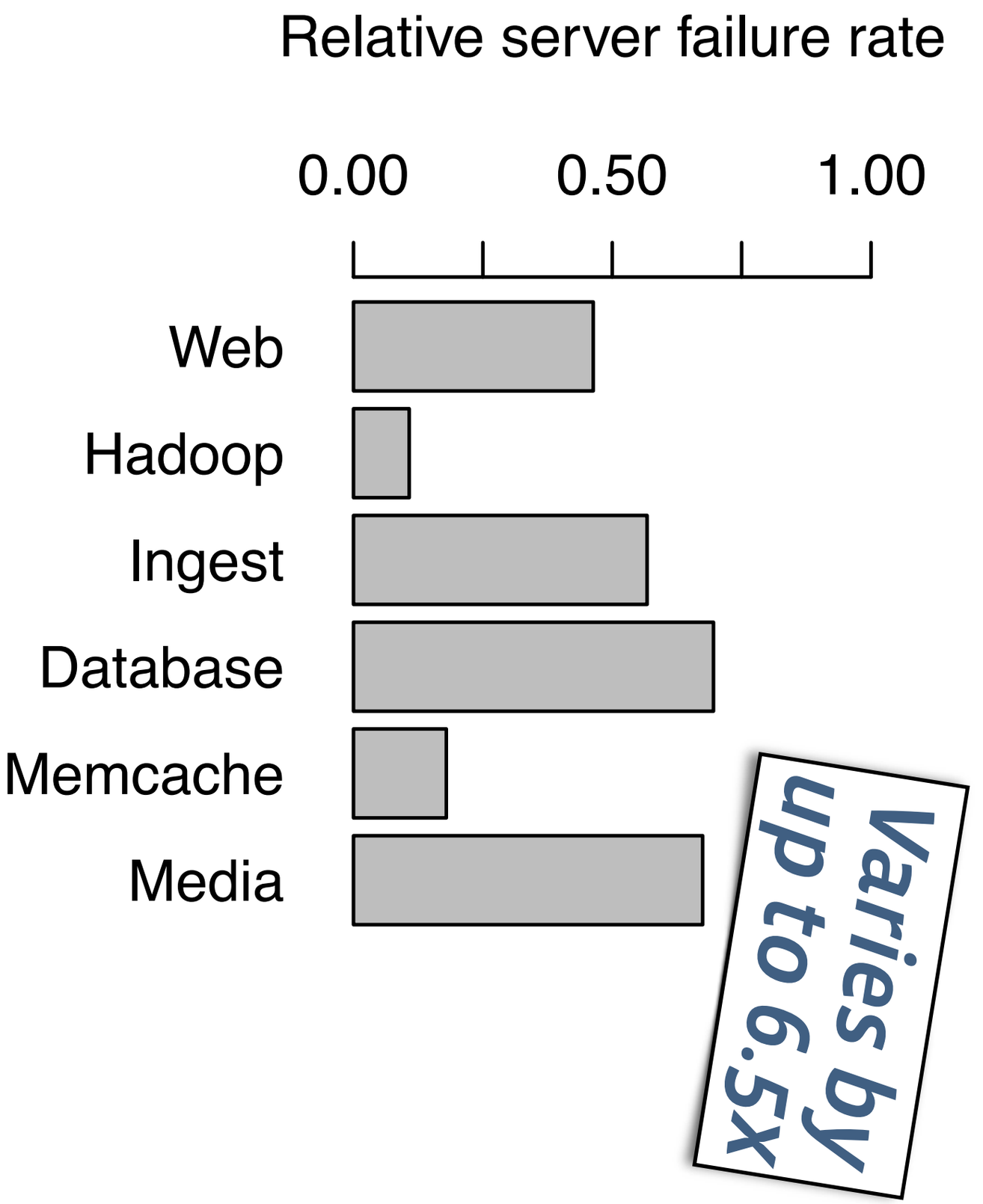
0.00

0 0.25 0.5 0.75 1

Memory utilization

No consistent trend across CPU/memory utilization


Relative server failure rate
0.00
0.50
1.00
Web
Hadoop
Ingest
Database
Memcache
Media


Relative server failure rate
0.00
0.50
1.00
Web
Hadoop
Ingest
Database
Memcache
Media
Varies by up to 6.5x

Error/failure occurrence

Page c *ogy*

Chips per DIMM, ***transfer width***, and ***workload type*** (not necessarily CPU/ memory utilization) affect reliability

trends

Modeling errors

Architecture & workload

Error/failure occurrence

Page offlining at scale

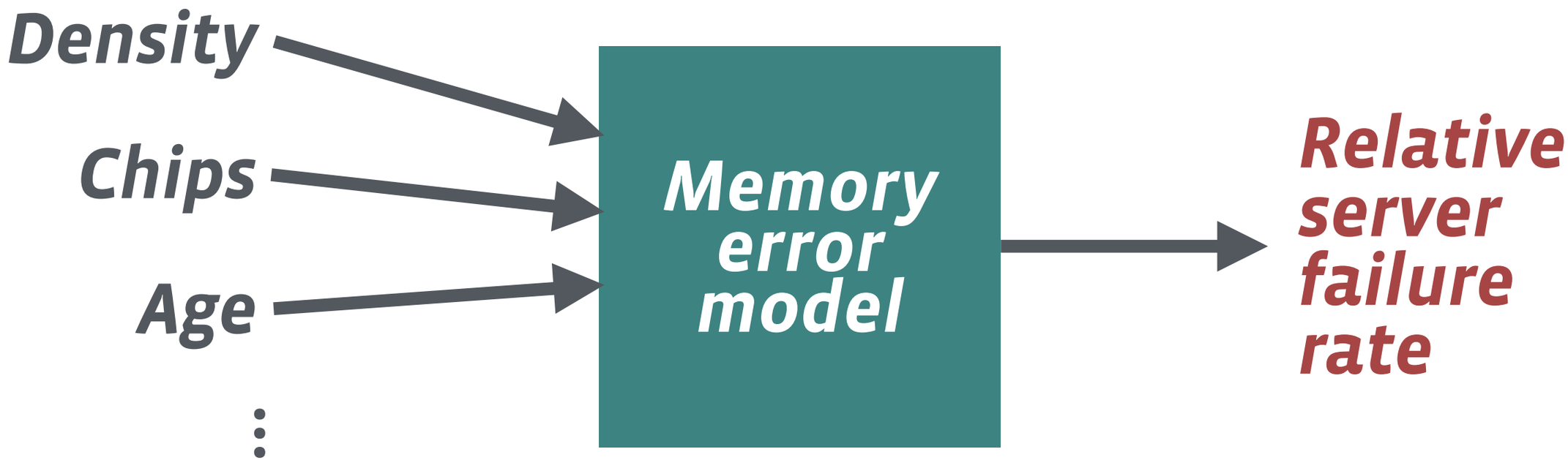
Technology scaling

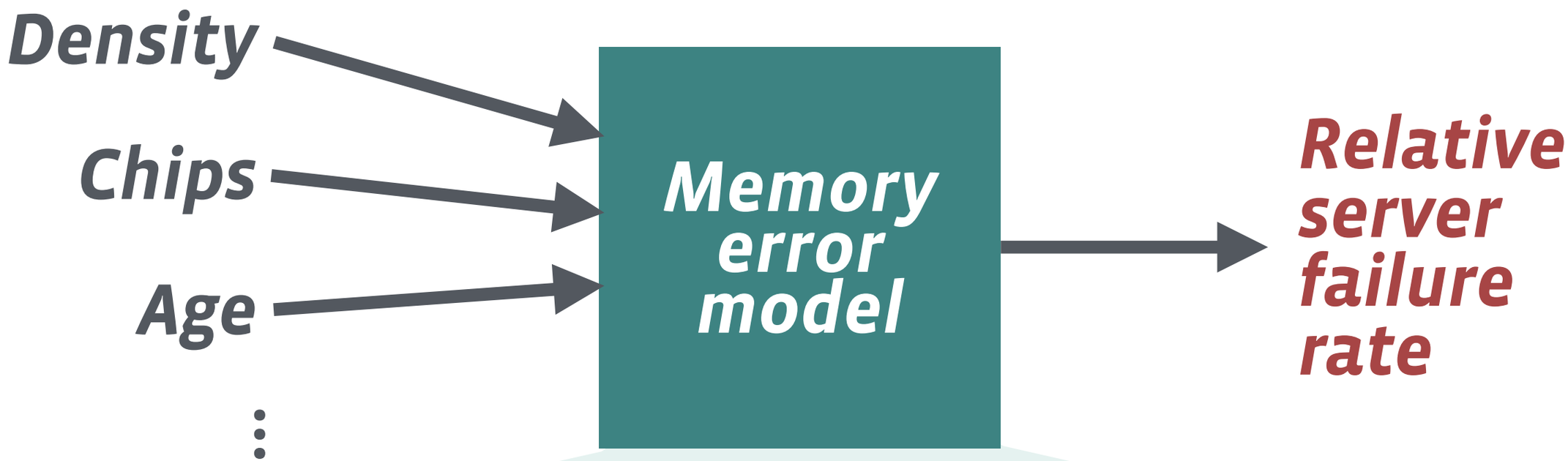
Modeling errors

Architecture & workload

A model for server failure

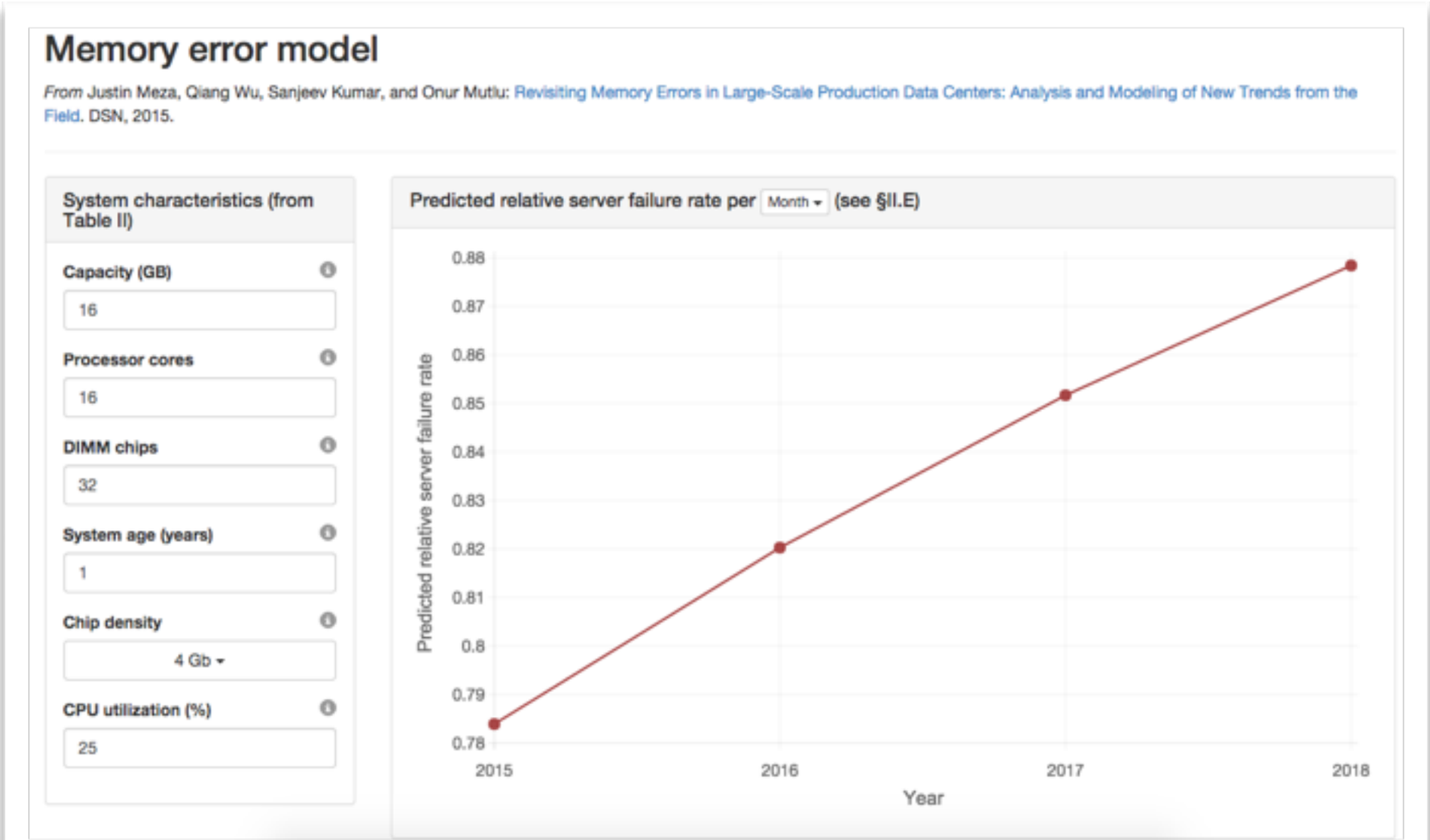
- use ***statistical regression model***
  - compare ***control group*** vs. ***error group***
  - ***linear regression*** in R
    - trained using data from analysis
- enable ***exploratory analysis***
  - high perf. vs. low power systems


Density
Chips
Age
Memory error model
Relative server failure rate



$$\ln[\mathcal{F}/(1-\mathcal{F})] = \beta_{Intercept} + (Capacity \cdot \beta_{Capacity}) + (Density2Gb \cdot \beta_{Density2Gb}) + (Density4Gb \cdot \beta_{Density4Gb}) + (Chips \cdot \beta_{Chips}) + (CPU\% \cdot \beta_{CPU\%}) + (Age \cdot \beta_{Age}) + (CPUs \cdot \beta_{CPUs})$$

Available online



http://www.ece.cmu.edu/~safari/tools/memerr/

Error/failure occurrence

Page c *ogy*

We have made publicly available a ***statistical model*** for assessing server memory reliability

trends

Modeling errors

Architecture & workload

Error/failure occurrence

Page offlining at scale

New reliability trends

Technology scaling

Modeling errors

Architecture & workload

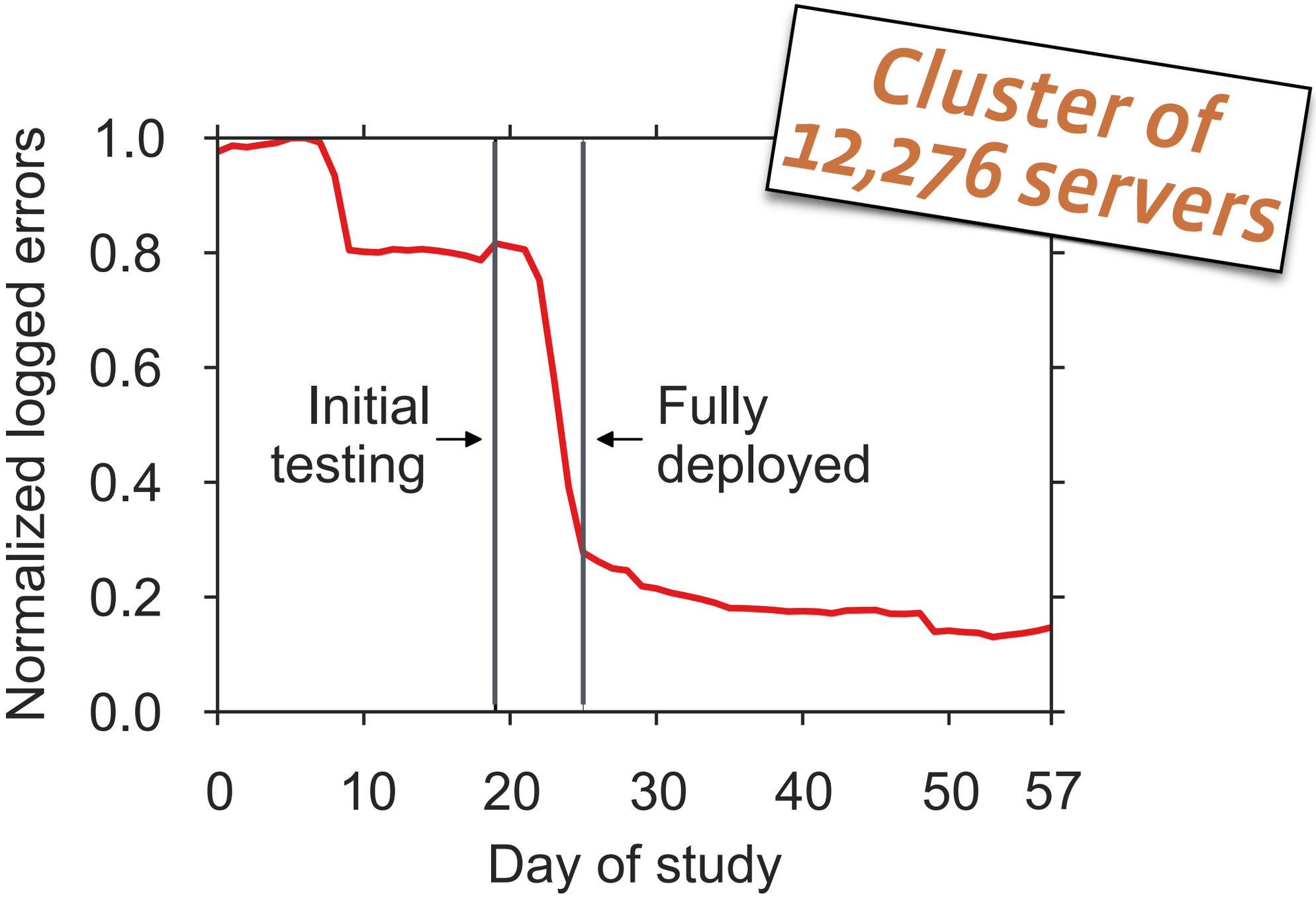
Prior page offlining work

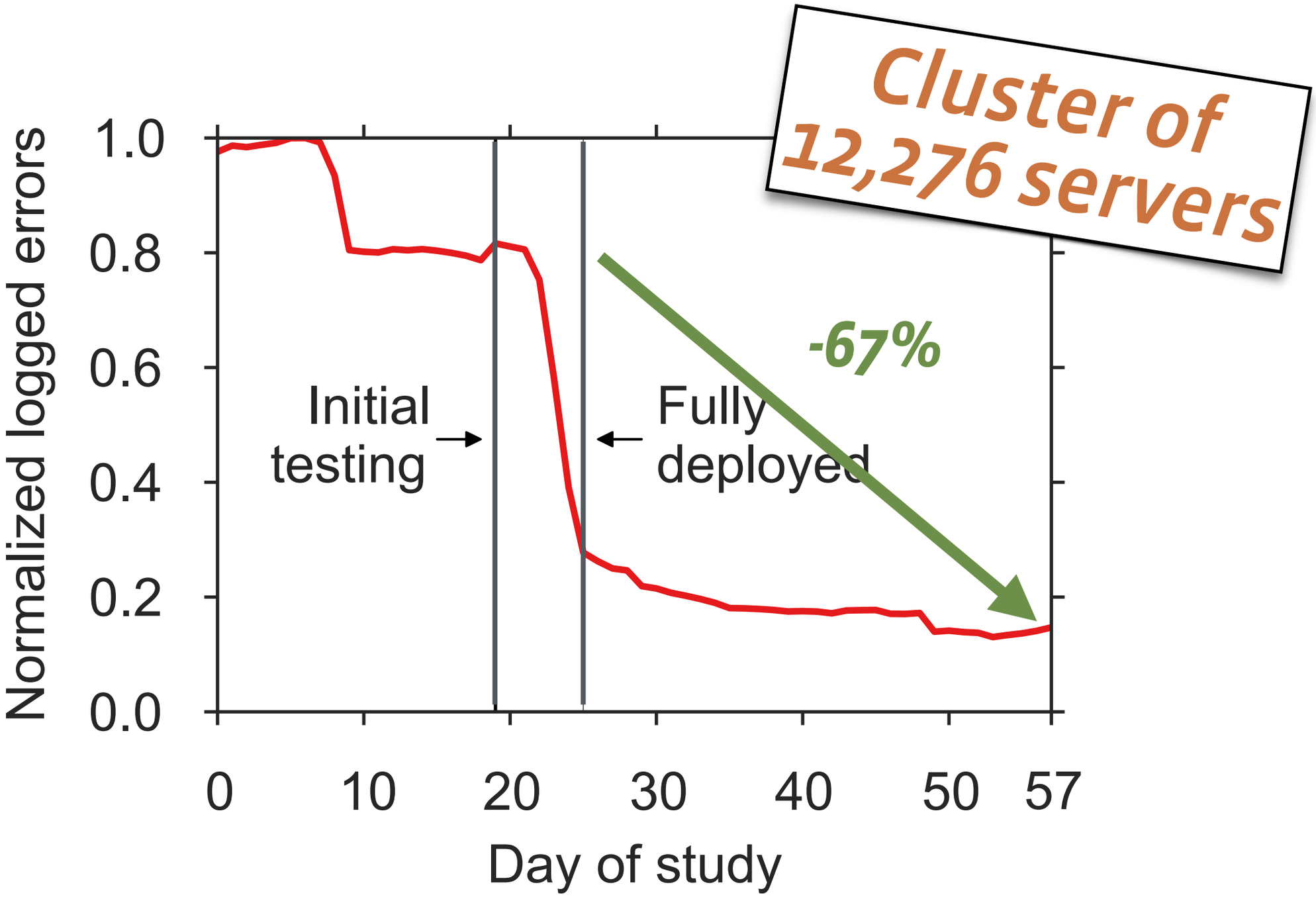
- [Tang+,DSN'06] proposed technique
  - "retire" faulty pages using OS
  - do not allow software to allocate them
- [Hwang+,ASPLOS'12] simulated eval.
  - error traces from Google and IBM
  - recommended retirement on first error
    - large number of cell/spurious errors

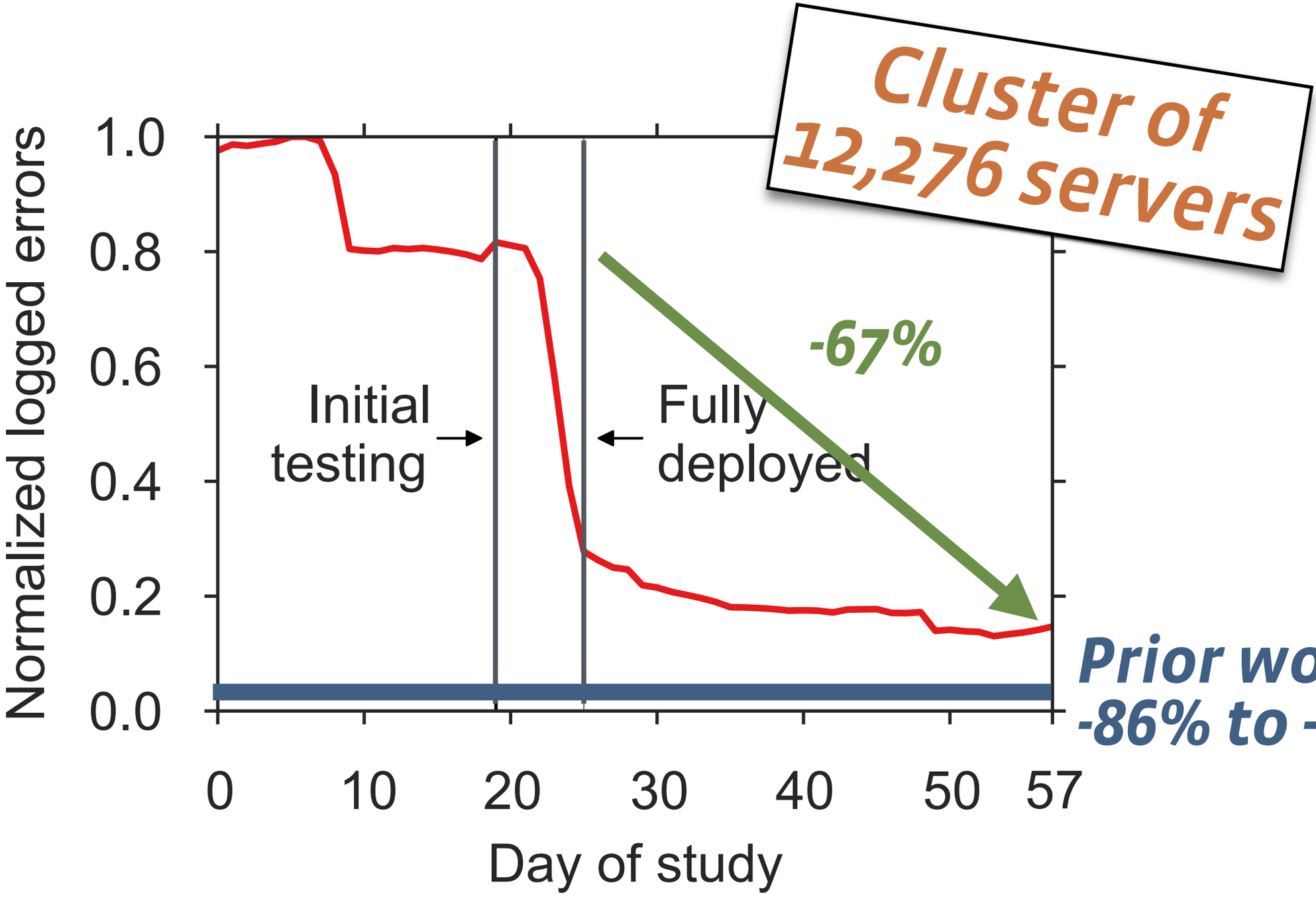
Prior page off lining work

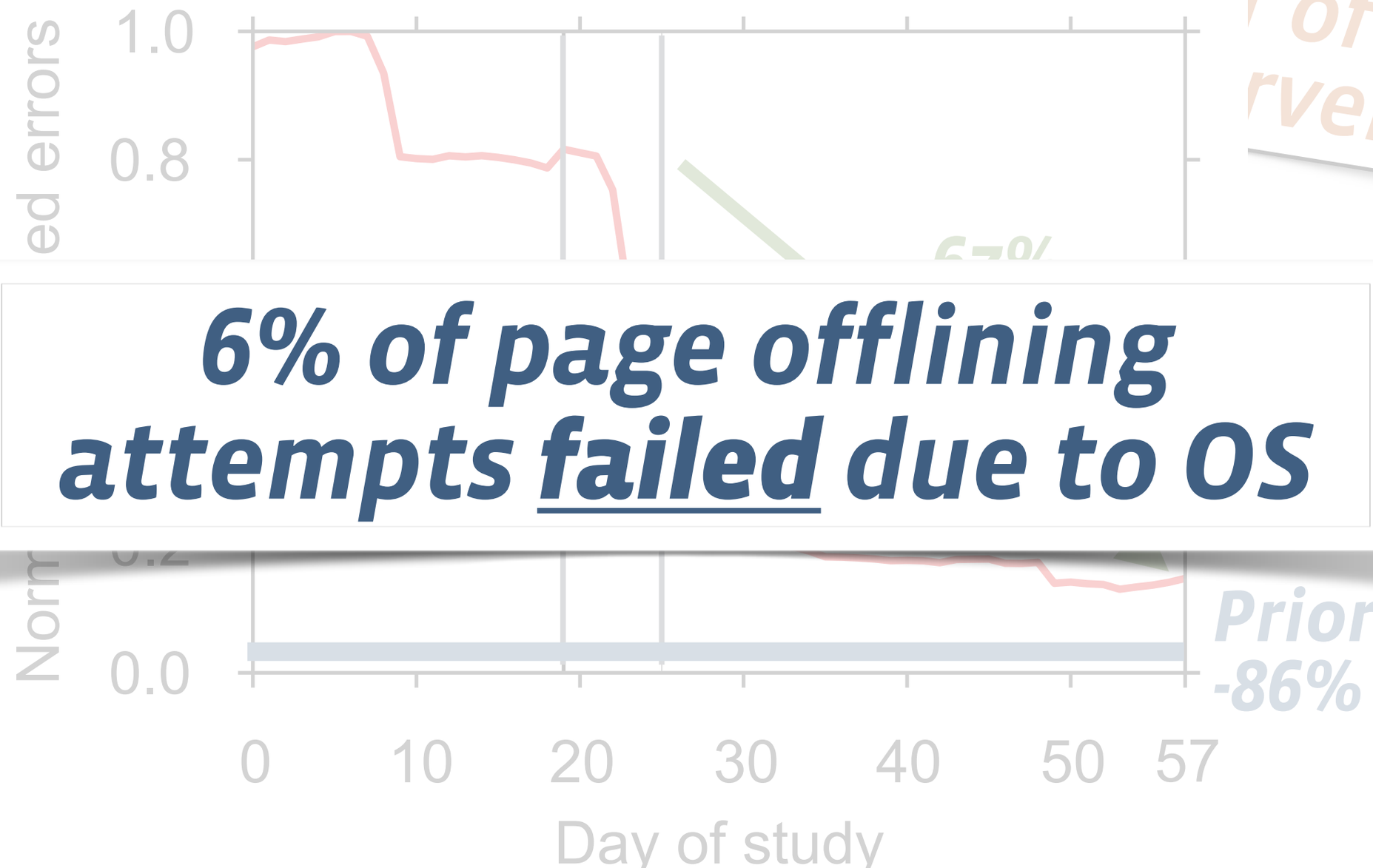
- [Tang+,DSN'06] proposed technique
  - "retire" faulty pages using OS
  - [illegible]
  - error traces from Google and IBM
  - recommended retirement on first error
    - large number of cell/spurious errors

How effective is page offlining in the wild?


Cluster of
12,276 servers
Normalized logged errors
1.0
0.8
0.6
0.4
0.2
0.0
Initial
testing
Fully
deployed
0
10
20
30
40
50
57
Day of study


Cluster of 12,276 servers
Normalized logged errors
1.0
0.8
0.6
0.4
0.2
0.0
Initial testing
Fully deployed
-67%
0
10
20
30
40
50
57
Day of study


Cluster of 12,276 servers
Normalized logged errors
1.0
0.8
0.6
0.4
0.2
0.0
Initial testing
Fully deployed
-67%
Prior work: -86% to -94%
0
10
20
30
40
50
57
Day of study



Cluster of servers

6-7%

Prior work:
-86% to -94%

6% of page offlining attempts failed due to OS

Error/failure occurrence

Page offlining at scale

First large-scale study of page offlining; real-world ***limitations*** of technique

trends

Modeling errors

Architecture & workload

Error/failure occurrence

Page offlining at scale

New reliability trends

Technology scaling

Modeling errors

Architecture & workload

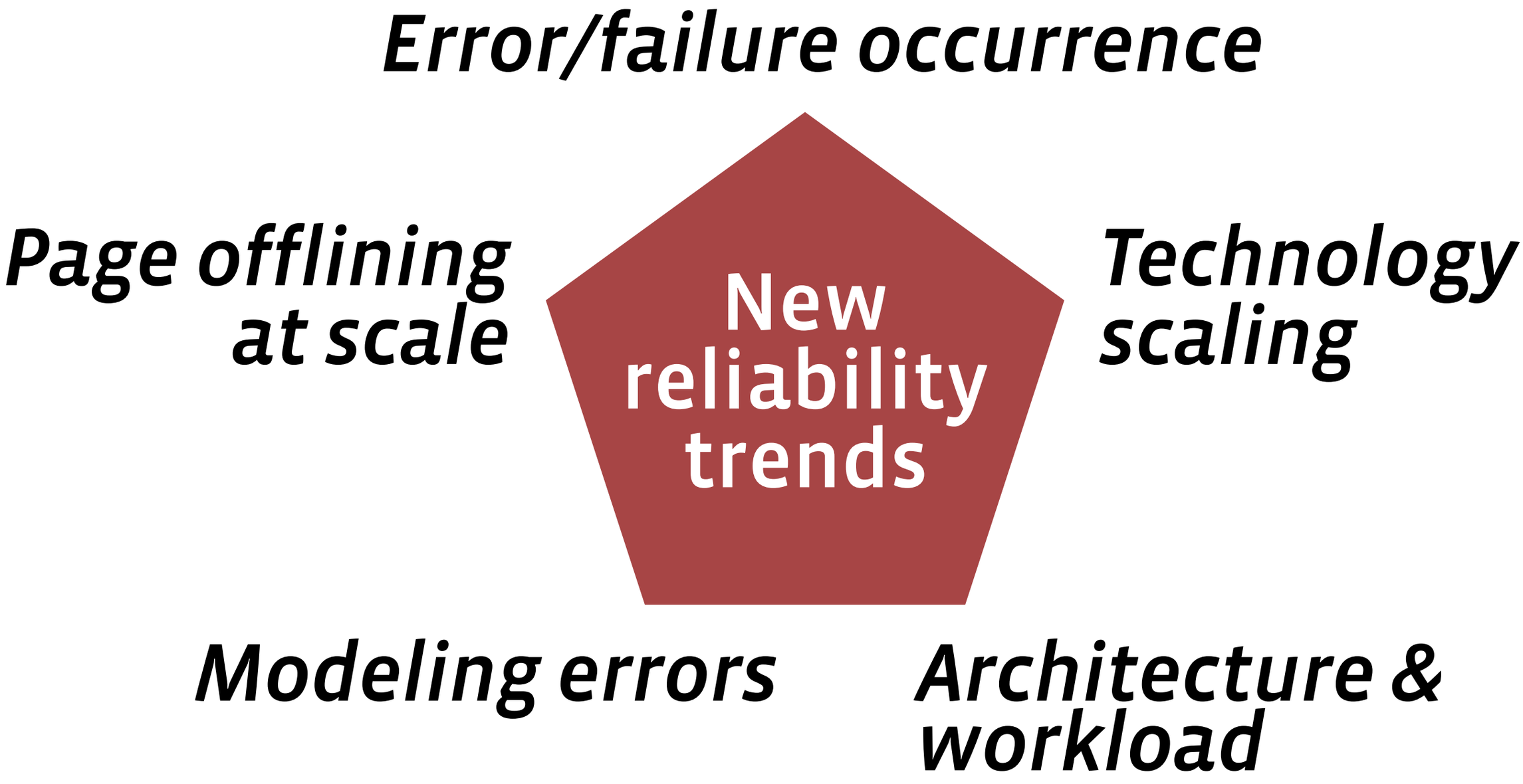
More results in paper

- *Vendors*
- *Age*
- *Processor cores*
- *Correlation analysis*
- *Memory model case study*

Summary

- *Modern systems*
- *Large scale*

Summary


Error/failure occurrence
Page offlining at scale
New reliability trends
Technology scaling
Modeling errors
Architecture & workload

Summary

Error/failure occurrence

Errors follow a ***power-law distribution*** and a large number of errors occur due to ***sockets/channels***

Modeling errors

Architecture & workload

Summary

Error/failure occurrence

We find that ***newer*** cell fabrication technologies have ***higher failure rates***

Technology scaling

trends

Modeling errors

Architecture & workload

Summary

Error/failure occurrence

Chips per DIMM, ***transfer width***, and ***workload type*** (not necessarily CPU/ memory utilization) affect reliability

Page c *ogy*

trends

Modeling errors

Architecture & workload

Summary

Error/failure occurrence

We have made publicly available a ***statistical model*** for assessing server memory reliability

Page c *ogy*

trends

Modeling errors

Architecture & workload

Summary

Error/failure occurrence

Page offlining at scale

First large-scale study of page offlining; real-world ***limitations*** of technique

trends

Modeling errors

Architecture & workload

Revisiting Memory Errors in Large-Scale Production Data Centers

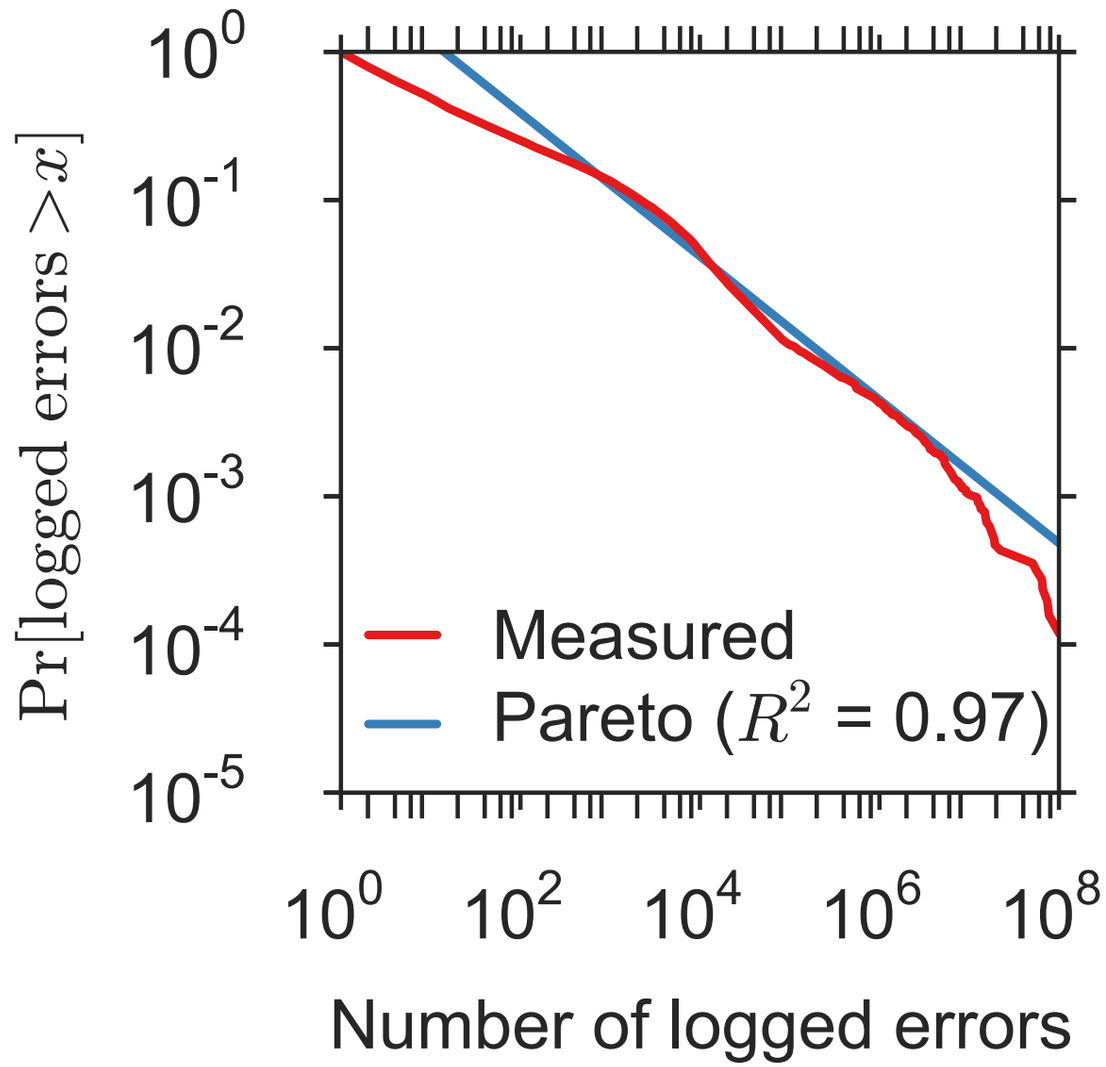
Analysis and Modeling of New Trends from the Field

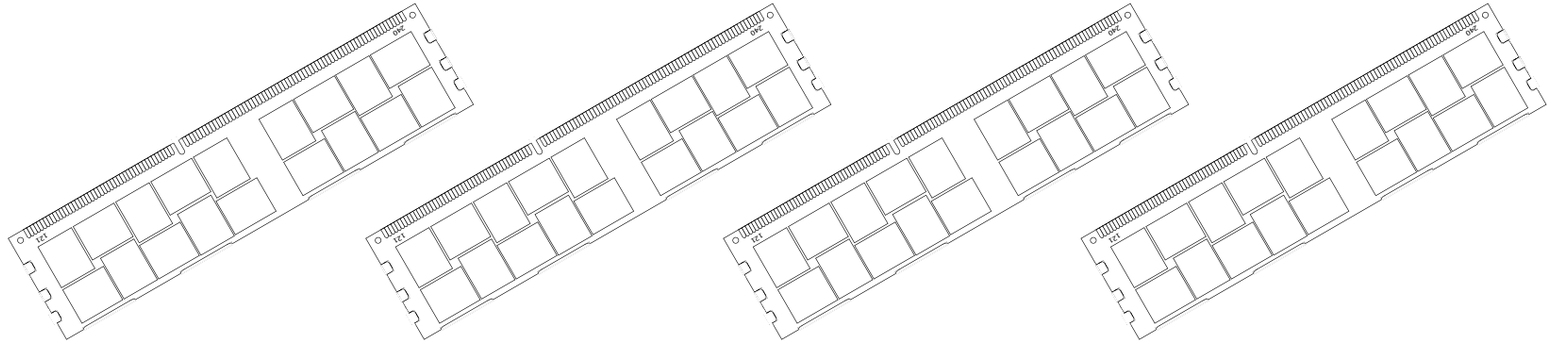
Justin Meza
Qiang Wu
Sanjeev Kumar
Onur Mutlu

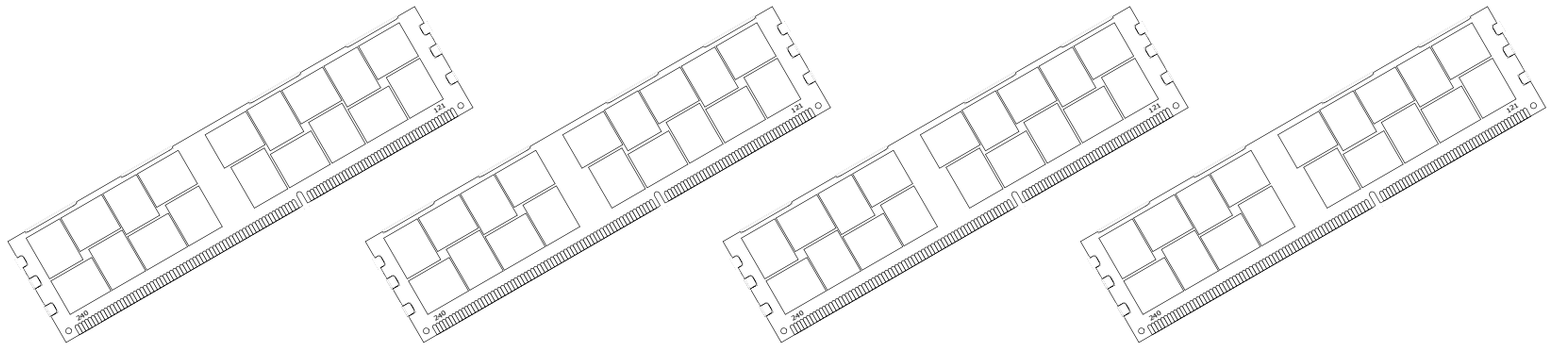
facebook
Carnegie Mellon University

Backup slides

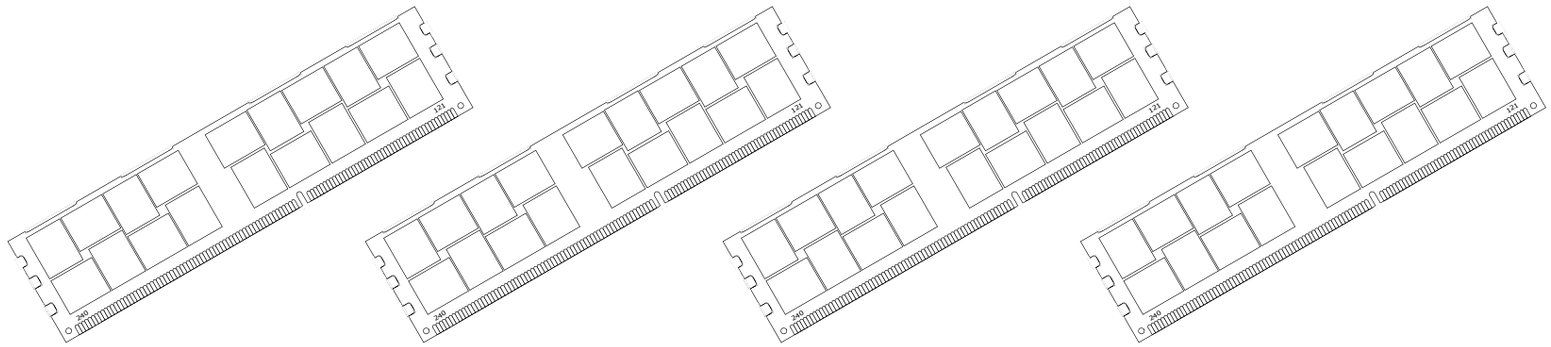
Decreasing hazard rate



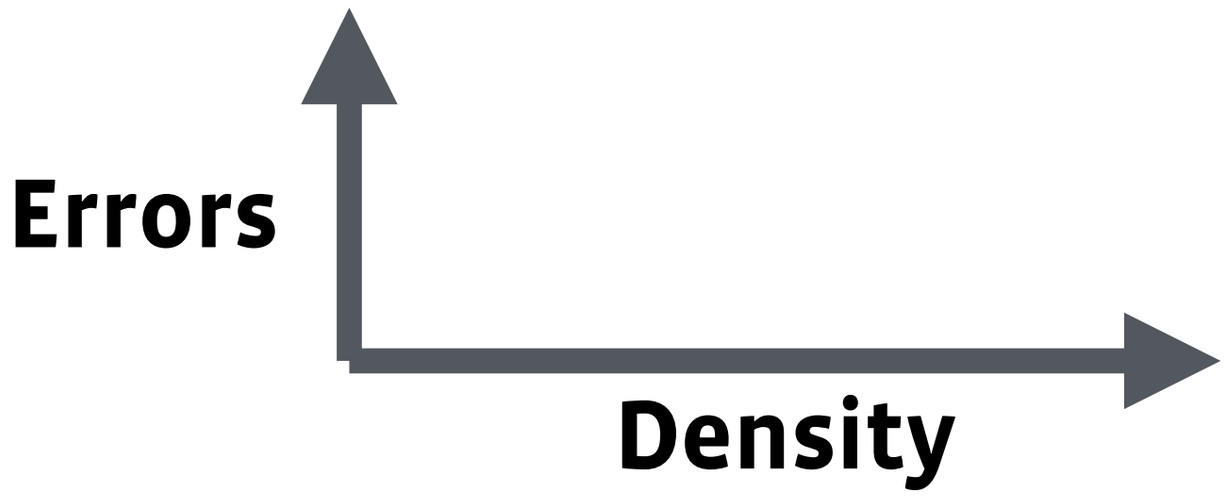

OUT
OUT
OUT
OUT

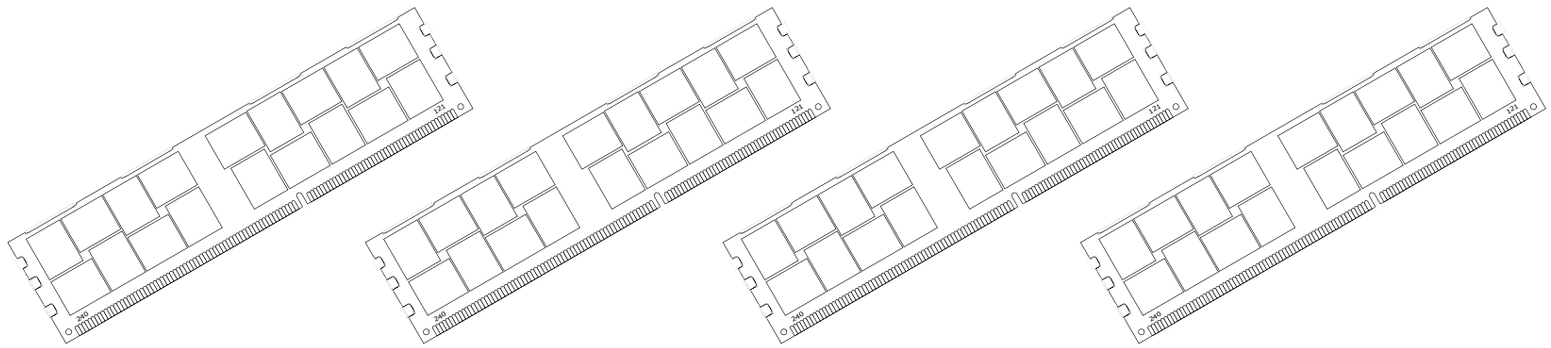


Errors	54,326	0	2	10
Density	4Gb	1Gb	2Gb	2Gb

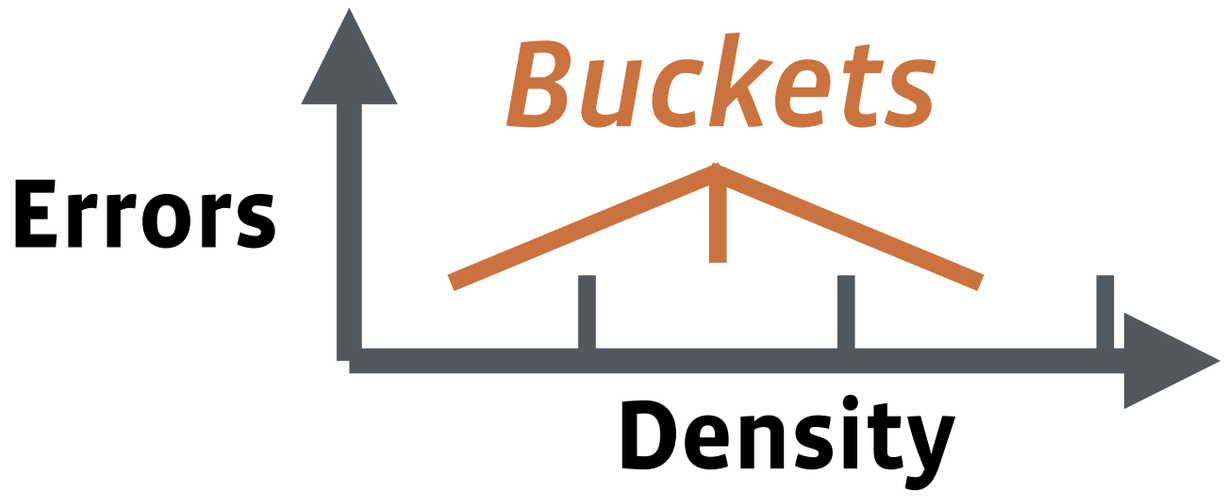


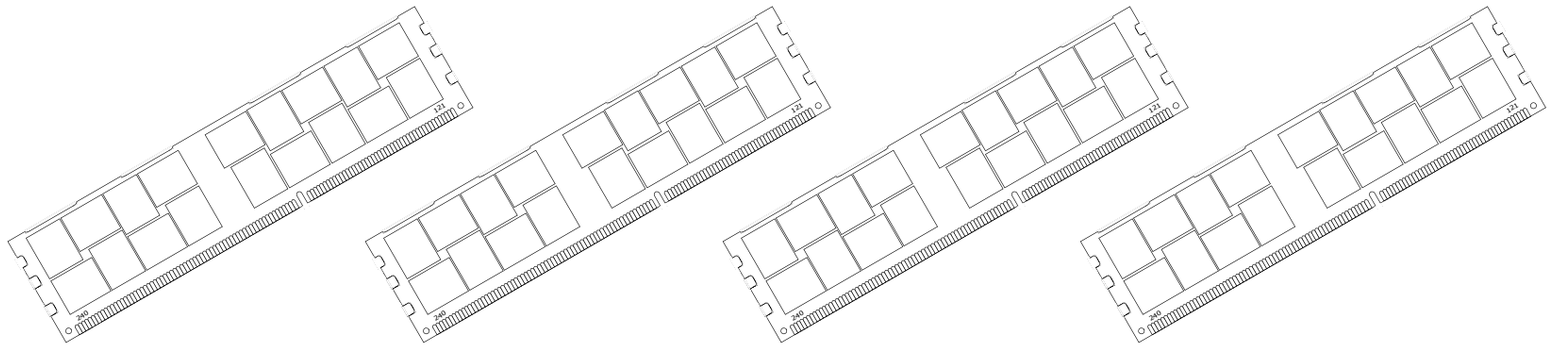
Errors	54,326	0	2	10
Density	4Gb	1Gb	2Gb	2Gb



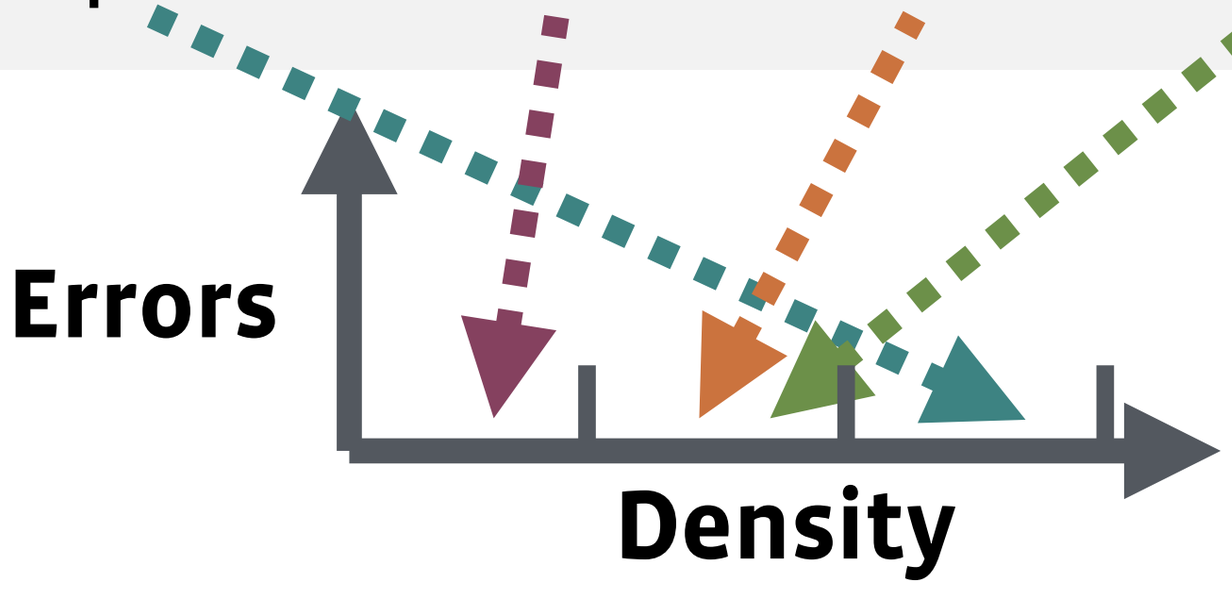


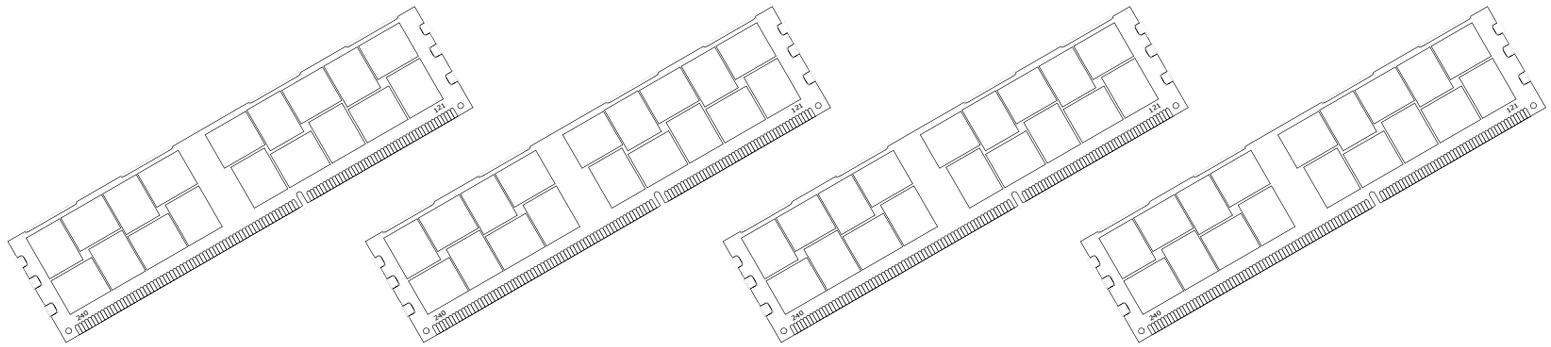
Errors	54,326	0	2	10
Density	4Gb	1Gb	2Gb	2Gb



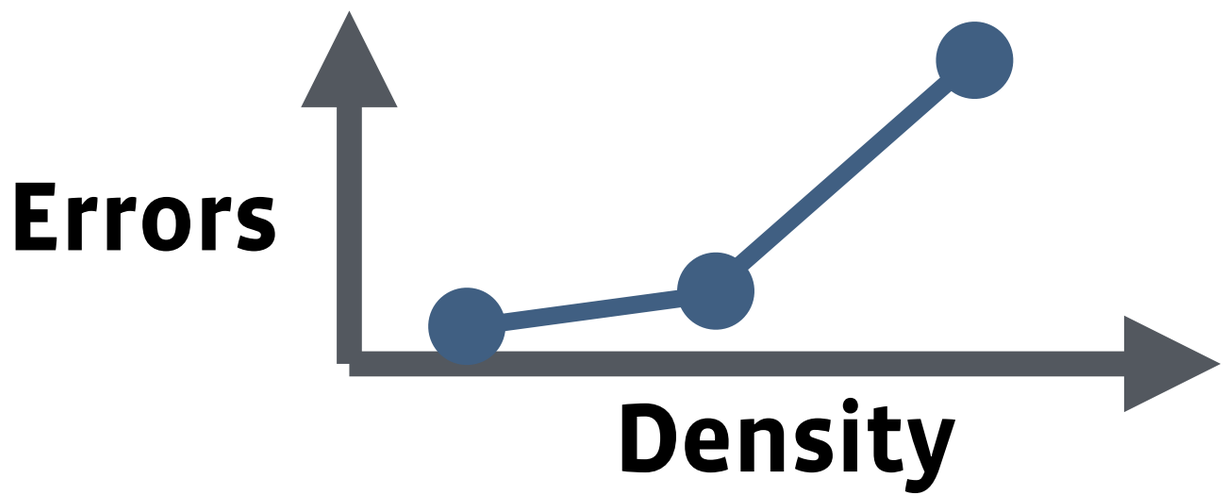


Errors	54,326	0	2	10
Density	4Gb	1Gb	2Gb	2Gb





Errors	54,326	0	2	10
Density	4Gb	1Gb	2Gb	2Gb



Case study

Case study

Factor	Low-end	High-end (HE)
Capacity	4 GB	16 GB
Density2Gb	1	0
Density4Gb	0	1
Chips	16	32
CPU%	50%	25%
Age	1	1
CPUs	8	16
Predicted relative failure rate	**0.12**	**0.78**

Inputs

Output

Case study

Factor	Low-end	High-end (HE)
Age	1	1
CPUs	8	16
Predicted relative failure rate	**0.12**	**0.78**

Does CPUs or density have a higher impact?

ts

Output

Exploratory analysis

Factor	Low-end	High-end (HE)	HE/↓density	HE/↓CPUs
Capacity	4 GB	16 GB	4 GB	16 GB
Density2Gb	1	0	1	0
Density4Gb	0	1	0	1
Chips	16	32	16	32
CPU%	50%	25%	25%	50%
Age	1	1	1	1
CPUs	8	16	16	8
Predicted relative failure rate	**0.12**	**0.78**	**0.33**	**0.51**

Exploratory analysis

Factor	Low-end	High-end (HE)	HE/↓density	HE/↓CPUs
Capacity	4 GB	16 GB	4 GB	16 GB
Density2Gb	1	0	1	0
Density4Gb	0	1	0	1
Chips	16	32	16	32
CPU%	50%	25%	25%	50%
Age	1	1	1	1
CPUs	8	16	16	8
Predicted relative failure rate	**0.12**	**0.78**	**0.33**	**0.51**