Baskin Engineering | UC SANTA CRUZ

Scale-out Edge Storage Systems with Embedded Storage Nodes to Get Better Availability and Cost-Efficiency At the Same Time

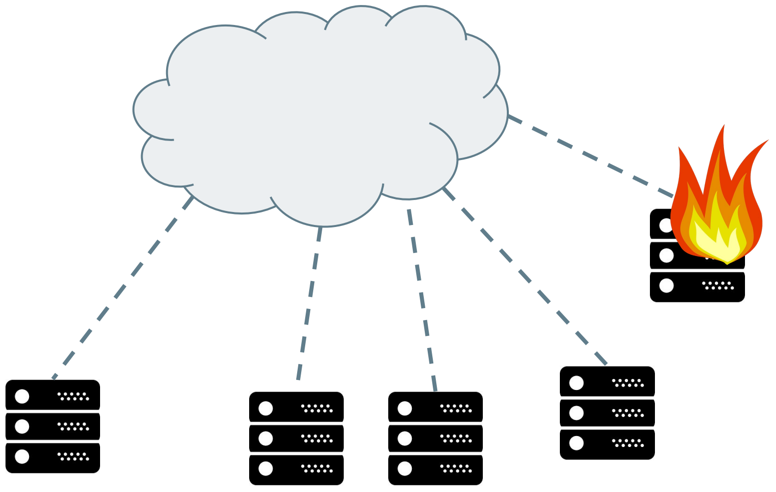
(aka "Embedded Storage at the Edge" Paper)

Jianshen Liu[*], Matthew Leon Curry[‡], Carlos Maltzahn[*], Philip Kufeldt[§]

[*]UC Santa Cruz, [‡]Sandia National Laboratories, [§]Seagate Technology

CROSS | CENTER FOR RESEARCH IN OPEN SOURCE SOFTWARE

Challenges of Data Availability at the Edge

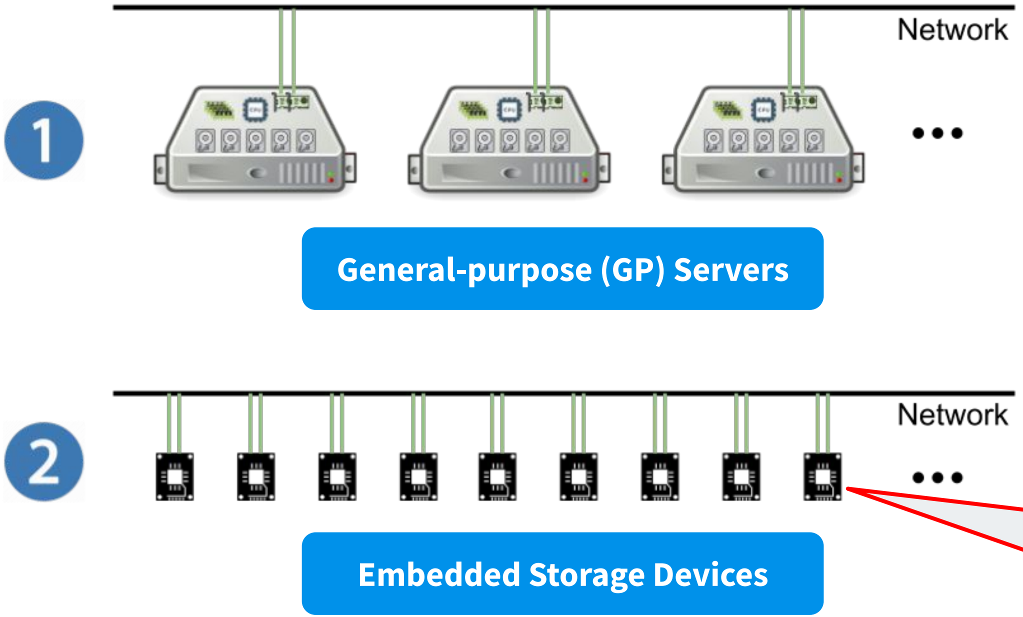


Edge Deployments

“Truck rolls” are expensive!

Environmental Limitations

Embedded Storage

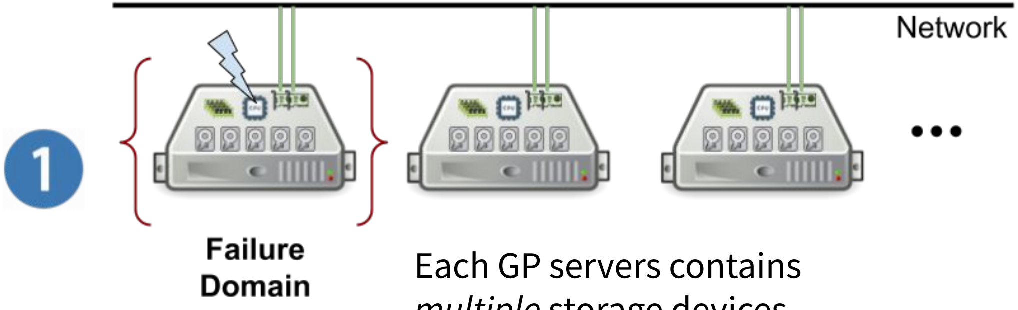


An Ethernet SSD with NVMe-oF Interface *

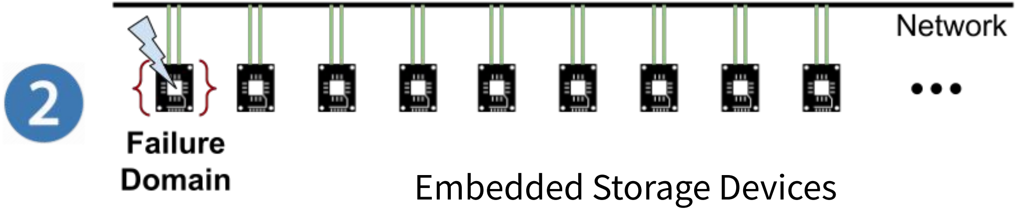
- ✓ Ethernet-attached storage devices integrated with computing resources
- ✓ Computational storage devices

* https://www.servethehome.com/marvell-88ss5000-nvmeof-ssd-controller-shown-with-toshiba-bics/

Failure Domains and Data Availability



Each GP servers contains *multiple* storage devices



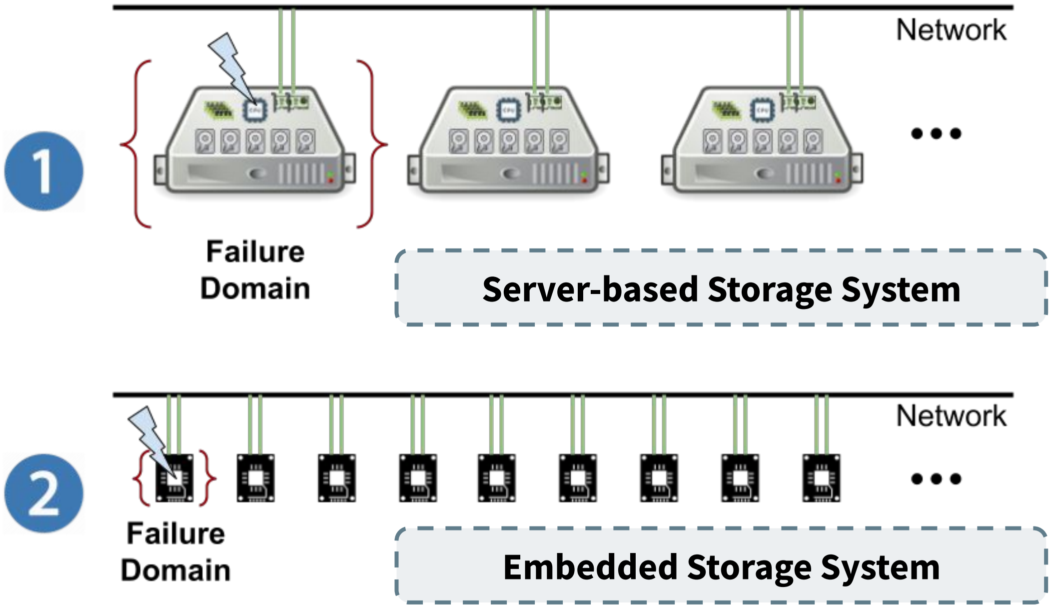
Embedded Storage Devices

Simpler

Embedded Storage enables **more nodes *under the same cost/space/power restrictions***.

The more independent failure domains a failover mechanism spans, the more available the data becomes.

The Analytical Model



Goal

Determine availability of embedded storage relative to traditional servers.

$$\text{Relative Benefit} = \frac{P_{\text{data-loss}}(\text{server-based storage system})}{P_{\text{data-loss}}(\text{embedded storage system})}$$

Relative Benefit > 1 → embedded storage is **better**

Our Analytical Model — Assumptions of System Configurations

- The units of deployment are homogeneous.
- Both systems have the same level of network redundancy and power redundancy for all nodes.
- Both systems use 3-way replication for data protection.
- Both systems use the **copyset replication**[§] scheme instead of the random replication scheme.

It's not our work, but we apply this scheme to our model

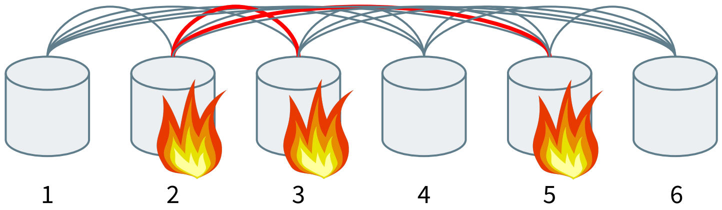
- Independence of servers and storage devices. Therefore, we can use *Poisson distribution*[*] to model the possibilities of hardware failures.

§ Cidon, Asaf, et al. "Copysets: Reducing the frequency of data loss in cloud storage." Presented as part of the 2013 {USENIX} Annual Technical Conference ({USENIX}{ATC} 13). 2013.
* Wikipedia contributors. "Poisson distribution." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 10 Mar. 2020. Web. 31 Mar. 2020.

Copyset Replication vs. Random Replication

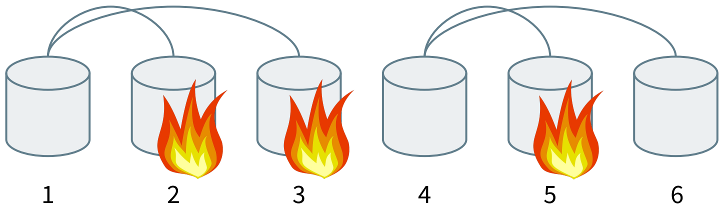
Replication Factor **r = 3**


: a node can store copies of the data in the other node



Relationships of Nodes with Random Replication

A node has replica set relationships with 5 nodes



Relationships of Nodes with Copyset Replication

A node has replica set relationships with <=2 nodes

With a sufficient number of data chunks stored, **data loss is nearly guaranteed if any combination of r nodes fail simultaneously.**

Reducing the number of replica sets can **reduce the likelihood of data loss under a correlated failure.**

Our Analytical Model — Assumptions of Model Parameters

Table 1: List of Model Parameters

Name	Description
m	the number of servers in the storage system
m'	the number of embedded storage devices in the storage system
n	the number of storage devices in a server
R_m	the failure rate of a server excluding the storage components
R_d	the failure rate of a block storage device in a server
R'_m	the failure rate of an embedded storage device excluding the storage component
R'_d	the failure rate of the storage component in an embedded storage device
w	the scatter width of the copyset replication

We use "m" to stands for "machine" and "d" for "device" in the notations of R_m, R_d, R'_m, and R'_d.

◎ $R_m = R'_m$ and $R_d = R'_d$

◎ $R_d = f \cdot R_m$, where $f > 0$

For hard drives, f could be greater than 2, while for SSDs, f could be less than 1.

(We call f **the ratio of failure rates**)

◎ $m' = c \cdot m$, where $c >= 1$

(We call c **the ratio of computing performance**)

◎ $n \geq 2$

(We call n **the ratio of storage performance**)

◎ $m \geq 3$ (3-way replication)

Our Analytical Model — Assumptions of Model Parameters

Table 1: List of Model Parameters

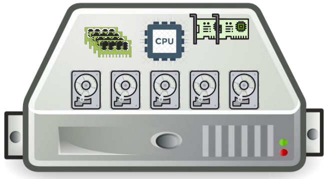
Name	Description
m	the number of servers in the storage system
m'	the number of embedded storage devices in the storage system
n	the number of storage devices in a server
R_m	the failure rate of a server excluding the storage components
R_d	the failure rate of a block storage device in a server
R'_m	the failure rate of an embedded storage device excluding the storage component
R'_d	the failure rate of the storage component in an embedded storage device
w	the scatter width of the copyset replication

We use "m" to stands for "machine" and "d" for "device" in the notations of R_m, R_d, R'_m, and R'_d.

◎ $R_m = R'_m$ and $R_d = R'_d$

Failure Rate of **non-storage components**

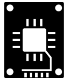
In



=

Failure Rate of **non-storage components**

In



Our Analytical Model — Assumptions of Model Parameters

Table 1: List of Model Parameters

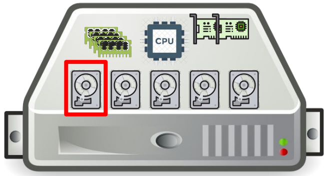
Name	Description
m	the number of servers in the storage system
m'	the number of embedded storage devices in the storage system
n	the number of storage devices in a server
R_m	the failure rate of a server excluding the storage components
R_d	the failure rate of a block storage device in a server
R'_m	the failure rate of an embedded storage device excluding the storage component
R'_d	the failure rate of the storage component in an embedded storage device
w	the scatter width of the copyset replication

We use "m" to stands for "machine" and "d" for "device" in the notations of R_m, R_d, R'_m, and R'_d.

◎ $R_m = R'_m$ and $R_d = R'_d$

Failure Rate of **a storage device**

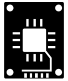
In



=

Failure Rate of **the storage component**

In



Our Analytical Model — Assumptions of Model Parameters

Table 1: List of Model Parameters

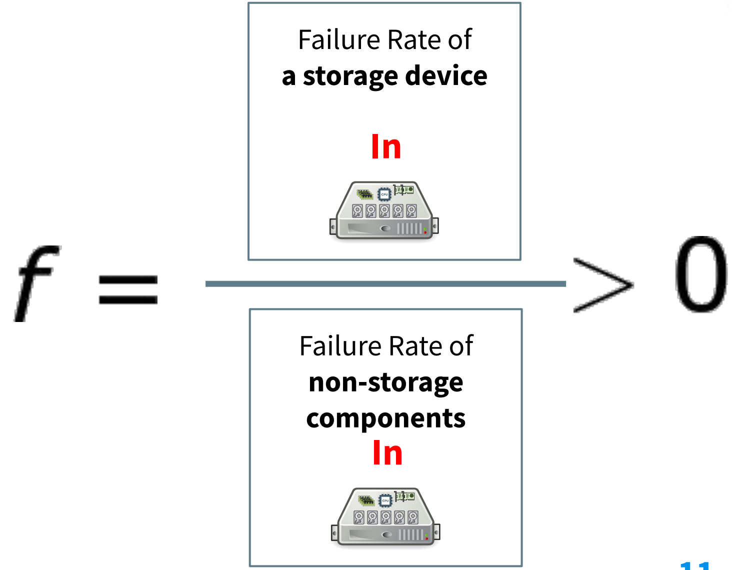
Name	Description
m	the number of servers in the storage system
m'	the number of embedded storage devices in the storage system
n	the number of storage devices in a server
R_m	the failure rate of a server excluding the storage components
R_d	the failure rate of a block storage device in a server
R'_m	the failure rate of an embedded storage device excluding the storage component
R'_d	the failure rate of the storage component in an embedded storage device
w	the scatter width of the copyset replication

We use "m" to stands for "machine" and "d" for "device" in the notations of R_m, R_d, R'_m, and R'_d.

◎ $R_d = f \cdot R_m$, where $f > 0$

For hard drives, f could be greater than 2, while for SSDs, f could be less than 1.

(We call f **the ratio of failure rates**)



Our Analytical Model — Assumptions of Model Parameters

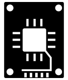
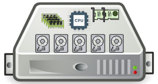
Table 1: List of Model Parameters

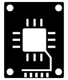
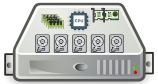
Name	Description
m	the number of servers in the storage system
m'	the number of embedded storage devices in the storage system
n	the number of storage devices in a server
R_m	the failure rate of a server excluding the storage components
R_d	the failure rate of a block storage device in a server
R'_m	the failure rate of an embedded storage device excluding the storage component
R'_d	the failure rate of the storage component in an embedded storage device
w	the scatter width of the copyset replication

We use "m" to stands for "machine" and "d" for "device" in the notations of R_m, R_d, R'_m, and R'_d.

◎ $m' = c \cdot m$, where $c \geq 1$

(We call c **the ratio of computing performance**)

$c =$ # of  / # of  ≥ 1

We need c units of  to get the same performance of a single 

Our Analytical Model — Assumptions of Model Parameters

Table 1: List of Model Parameters

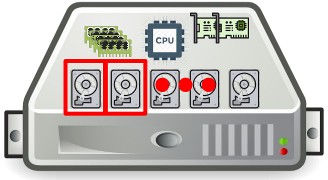
Name	Description
m	the number of servers in the storage system
m'	the number of embedded storage devices in the storage system
n	the number of storage devices in a server
R_m	the failure rate of a server excluding the storage components
R_d	the failure rate of a block storage device in a server
R'_m	the failure rate of an embedded storage device excluding the storage component
R'_d	the failure rate of the storage component in an embedded storage device
w	the scatter width of the copyset replication

We use "m" to stands for "machine" and "d" for "device" in the notations of R_m, R_d, R'_m, and R'_d.

◎ $n \geq 2$

(We call n **the ratio of storage performance**)

n is the number of storage devices (≥ 2) in a server.



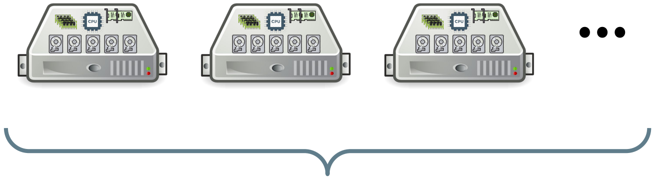
Our Analytical Model — Assumptions of Model Parameters

Table 1: List of Model Parameters

Name	Description
m	the number of servers in the storage system
$m^{'}$	the number of embedded storage devices in the storage system
n	the number of storage devices in a server
R_m	the failure rate of a server excluding the storage components
R_d	the failure rate of a block storage device in a server
$R_m^{'}$	the failure rate of an embedded storage device excluding the storage component
$R_d^{'}$	the failure rate of the storage component in an embedded storage device
w	the scatter width of the copyset replication

We use "m" to stands for "machine" and "d" for "device" in the notations of R_m, R_d, $R_m^{'}$, and $R_d^{'}$.

◎ $m \geq 3$ (3-way replication)



need at least 3 servers for 3-way replication

Our Analytical Model — Assumptions of Model Parameters

Table 1: List of Model Parameters

Name	Description
m	the number of servers in the storage system
m'	the number of embedded storage devices in the storage system
n	th
R_m	th c
R_d	th ser
R'_m	the failure rate of an embedded storage device excluding the storage component
R'_d	the failure rate of the storage component in an embedded storage device
w	the scatter width of the copyset replication

We use "m" to stands for "machine" and "d" for "device" in the notations of R_m, R_d, R'_m, and R'_d.

How sensitive is the Relative Benefit to these parameters?

- $R_m = R'_m$ and $R_d = R'_d$
- $R_d = f \cdot R_m$, where $f > 0$

  For hard drives, f could be greater than 2, while for SSDs, f could be less than 1.

  (We call f **the ratio of failure rates**)
- $m' = c \cdot m$, where $c >= 1$

  (We call c **the ratio of computing performance**)
- $n \geq 2$

  (We call n **the ratio of storage performance**)
- $m \geq 3$ (3-way replication)

Evaluation

As an example, we evaluate the **Relative Benefit** of embedded storage regarding the data unavailability caused by failures of exactly **three** components.

A component can be:

- A server
- An embedded storage device
- A storage component in a failure domain

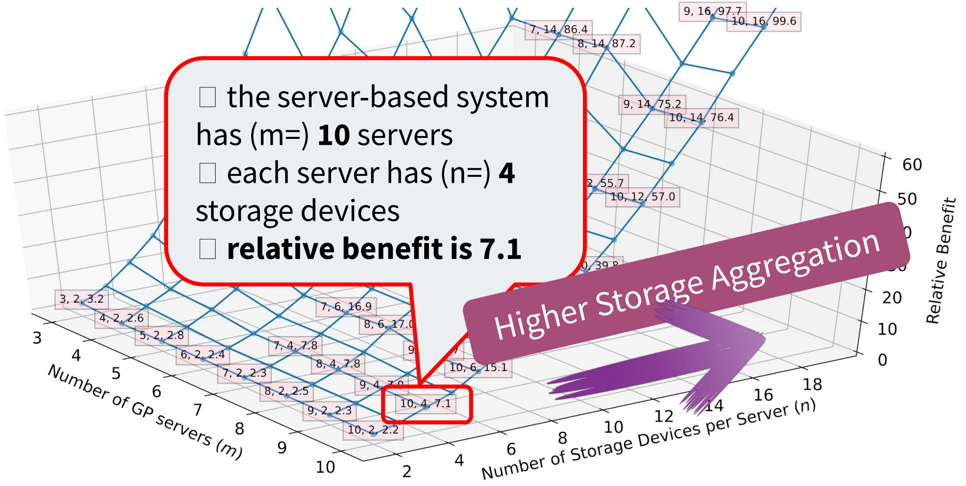
$$\textbf{Relative Benefit} = \frac{P_{\text{data-loss}}(\text{server-based storage system})}{P_{\text{data-loss}}(\text{embedded storage system})}$$

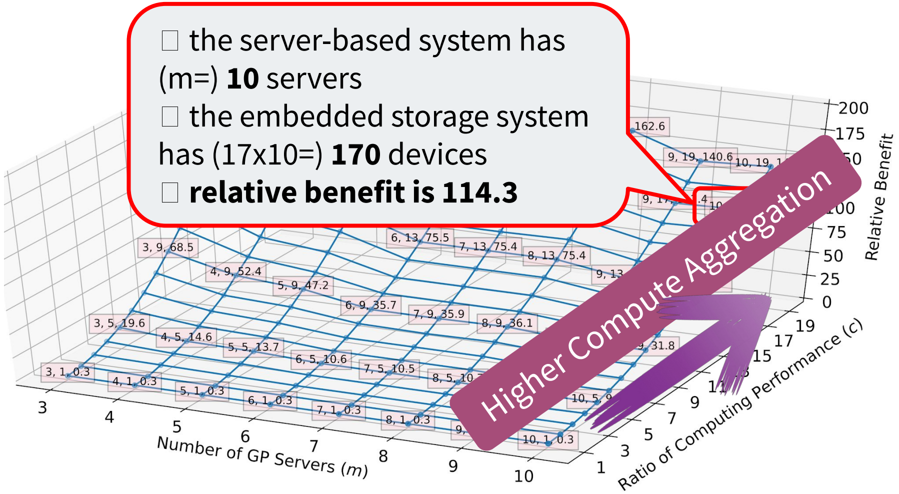
✓ f (the failure rate of the storage component over the failure rate of the non-storage components)
✓ w (the number of nodes that have a replica set relationship with a node)
➔ m (# of GP servers)
➔ n (# of storage devices in a server)
➔ c (# of embedded storage device / # of servers)

❓ $f_{\text{relative\_benefit}}(m, n)$ and ❓ $f_{\text{relative\_benefit}}(m, c)$

Evaluation – Spinning Media as Storage

- The failure rate of a storage device is **2x** of that of the non-storage components of a server (**f = 2**)
  [Vishwanath, et al. "Characterizing cloud computing hardware reliability." 2010]
- The number of nodes that have a replica set relationship with a node is 4 (**w = 4**)





The Impact of **Storage Aggregation** on the Relative Benefit

$(c = n)$

c = n = 4 ➡ the embedded storage system has (10 … devices

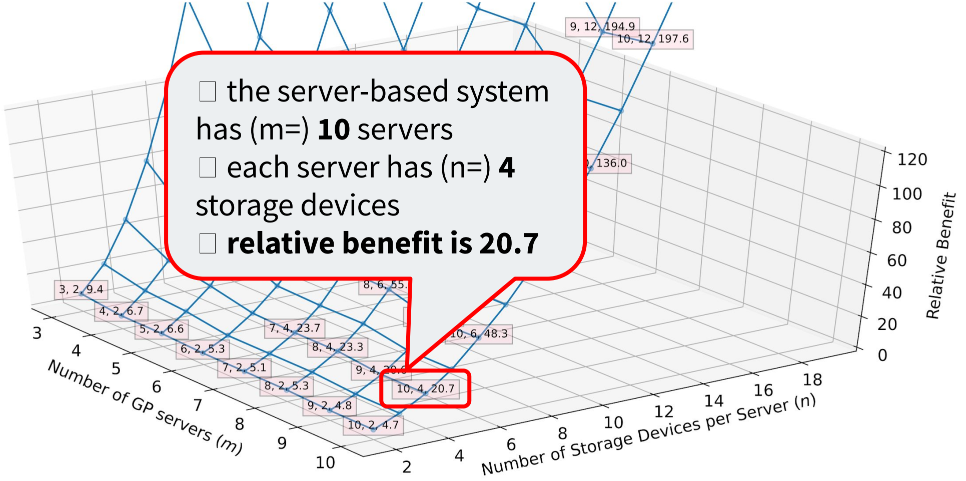
each server has **12** storage devices

$(n = 12)$

The Impact of **Compute Aggregation** on the Relative Benefit

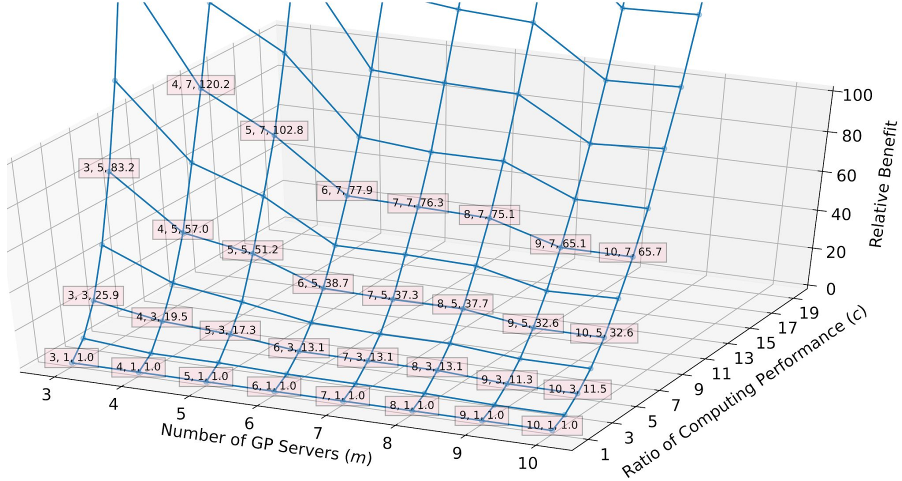
Evaluation – Solid-state Drives as Storage

- The failure rate of a storage device is **0.06x** of that of the non-storage components of a server (**f = 0.06**)
  [Xu, Erci, et al. "Lessons and actions: What we learned from 10k ssd-related storage system failures." 2019]
- The number of nodes that have a replica set relationship with a node is 4 (**w = 4**)



The Impact of **Storage Aggregation** on the Relative Benefit

$(c = n)$

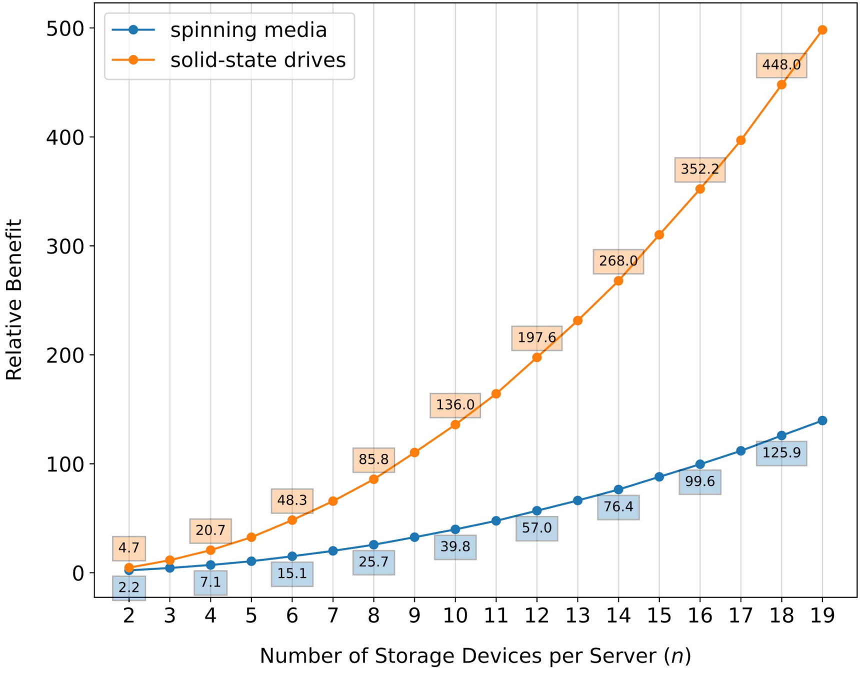


The Impact of **Compute Aggregation** on the Relative Benefit

$(n = 12)$

Insights (part 1/5)

1. The higher the storage aggregation of a server, the higher the relative benefit of embedded storage.



Server-based Storage System

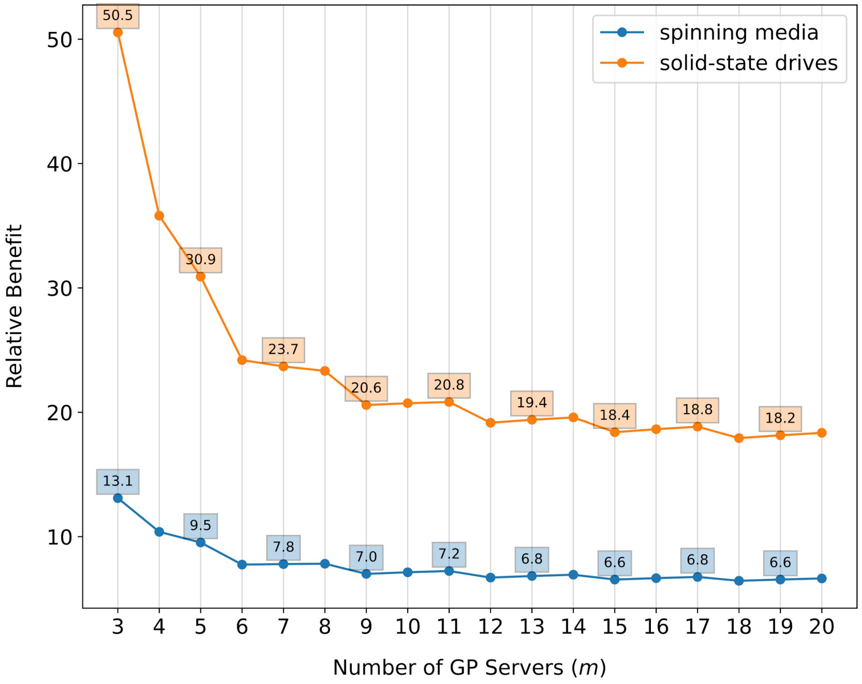
10 servers with **n** storage devices each, resulting in 10 failure domains.

Embedded Storage System

10 x **n** devices, resulting in 10 x **n** failure domains.

Insights (part 2/5)

2. Smaller storage systems are more sensitive to the benefit of embedded storage.



Server-based Storage System

m servers have 4 storage devices each, resulting in **m** failure domains.

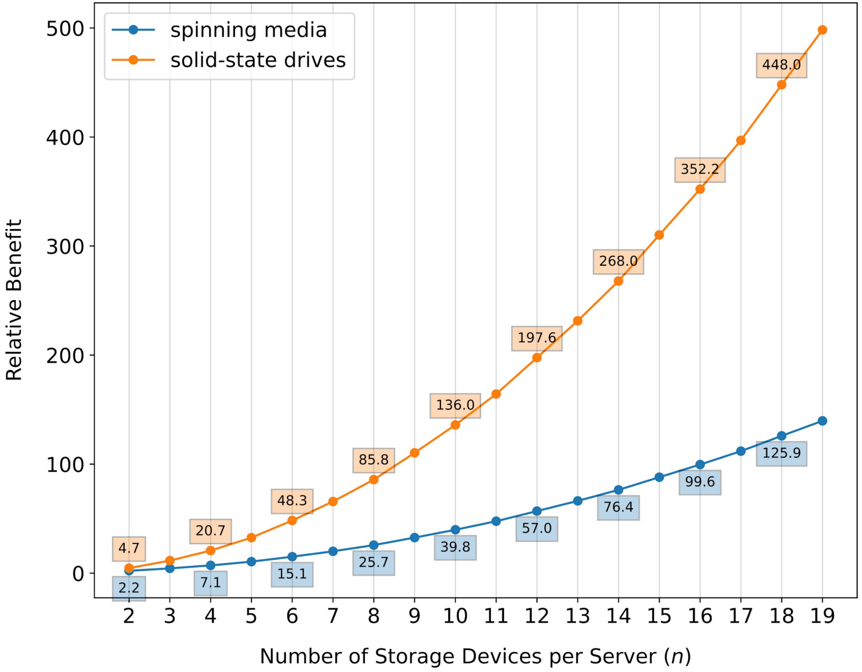
Embedded Storage System

4 x **m** devices,
resulting in 4 x **m** failure domains.

The total # of storage devices of the two systems are the same.

Insights (part 3/5)

3. The lower the failure rate of a storage device, the higher the relative benefit of embedded storage.



Server-based Storage System

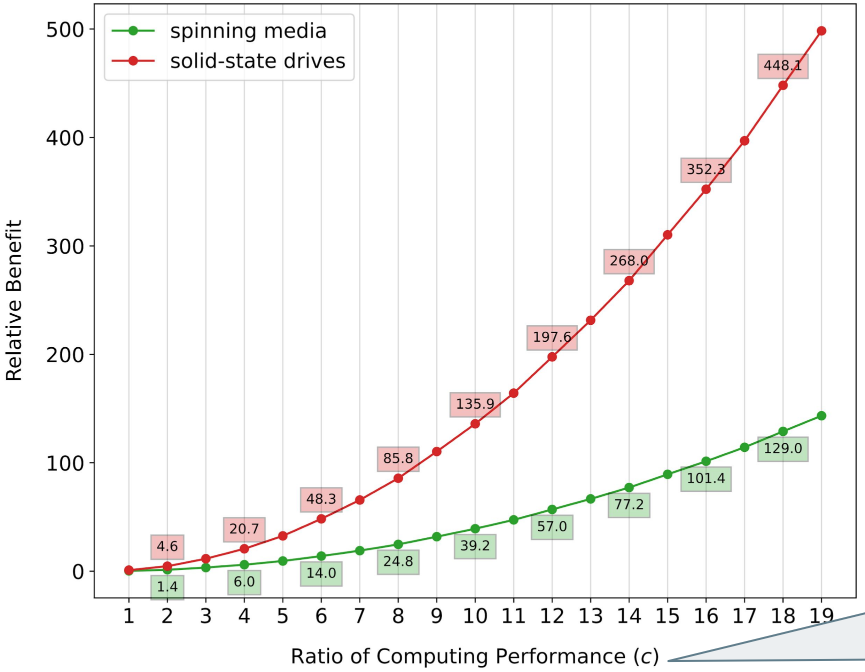
10 servers with **n** storage devices each, resulting in 10 failure domains.

Embedded Storage System

10 x **n** devices, resulting in 10 x **n** failure domains.

Insights (part 4/5)

4. The higher the compute aggregation of a server, the higher the relative benefit of embedded storage.

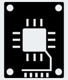
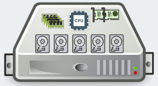


Server-based Storage System

10 servers with 12 storage devices each

Embedded Storage System

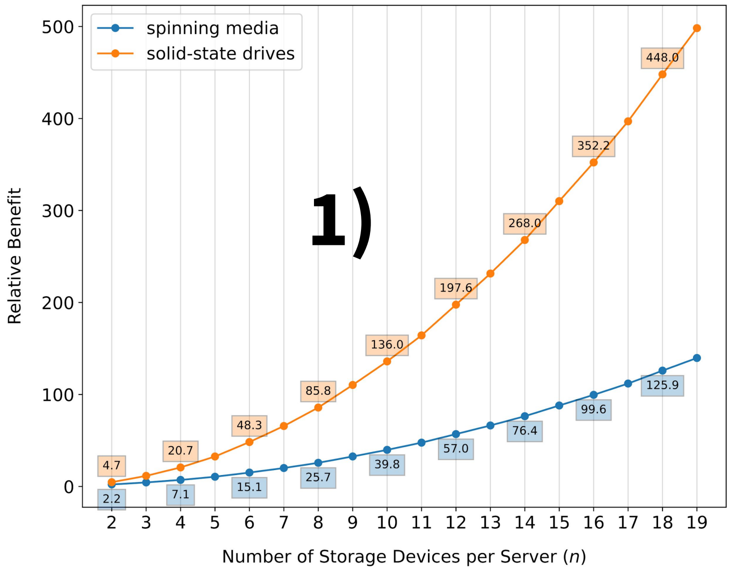
10 x **c** devices

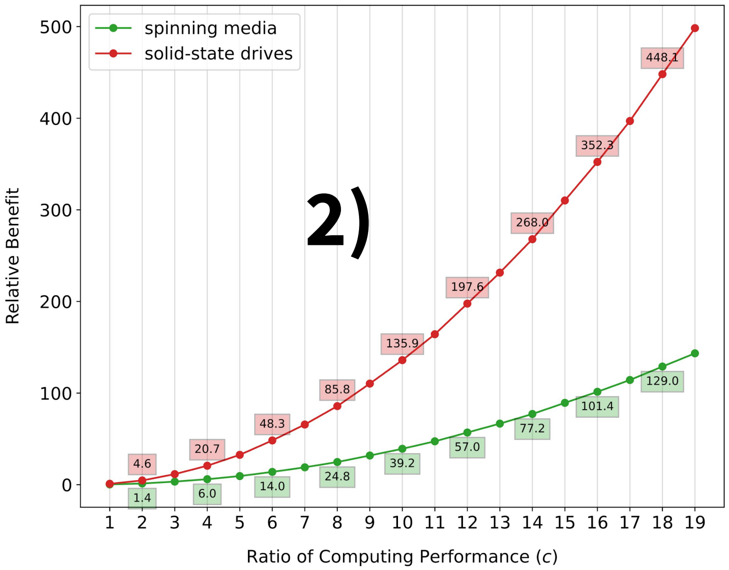
C units of  can provide the same storage performance of a single 

Insights (part 5/5)

5. The relationship between the resource aggregation and the relative benefit is nonlinear.

1) Doubling the storage aggregation of a server could triple the relative benefit.
2) Doubling the compute aggregation of a server could quadruple the relative benefit.





Conclusions

- Embedded storage devices are simpler, making it is possible to have more independent failure domains.

- Storage systems with more independent failure domains can improve data availability.

- A great design point, but many unsolved challenges!
(e.g., explore the balance between availability and storage performance)

Baskin Engineering UC SANTA CRUZ

This work was supported in part by NSF grants OAC-1836650, CNS-1764102, and CNS-1705021, and by the Center for Research in Open Source Software (cross.ucsc.edu). Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Thank you!

Questions?

Jianshen Liu

jliu120@ucsc.edu

https://cross.ucsc.edu (Eusocial Storage Devices)

CROSS | CENTER FOR RESEARCH IN OPEN SOURCE SOFTWARE

An Example of Copyset Replication

- A **copyset** is a set of nodes that stores all of the copies of a data chunk.
- **Scatter width** is the number of nodes the data of a node can be replicated to.
- Example:

# of nodes (m)	replication factor (r)	scatter width (w)
9	3	4

Copysets: {1,2,3}, {4,5,6}, {7,8,9}
{1,4,7}, {2,5,8}, {3,6,9}

$$\frac{w}{r-1} = 2 \quad \text{permutations}$$

- Each permutation increases the scatter width of a node by $r - 1$
- The number of copysets is $\frac{w}{r-1}\frac{m}{r}$

Copyset Replication vs. Random Replication

- Number of copysets (3-way replication):

Copyset Replication (CR)	Random Replication (RR)
$\frac{w}{r-1}\frac{m}{r} = \frac{wm}{6}$	$\binom{m}{3} = \frac{m(m-1)(m-2)}{6}$

$$\frac{\#\text{ of copysets using RR}}{\#\text{ of copysets using CR}} = \frac{(m-1)(m-2)}{w}$$

- With a sufficient number of data chunks stored, random replication creates a failure domain for **any combination of r nodes** (r is the replication factor).

Our Analytical Model — Modeling the Two Systems

The possibility of data loss of server-based storage systems

$$P(\text{failures of k servers}) = \frac{{R_m}^k e^{-R_m}}{k!}$$

$$P_{gp} = \sum_{k=3}^{m} P_m(k) + \sum_{j=3}^{mn} P_d(j) + \sum_{k=2}^{m}\sum_{j=1}^{mn} P_{m,d}(k,j) + \sum_{j=2}^{mn} P_{m,d}(1,j)$$

where

$$P_m(k) = P(\text{failures of } k \text{ servers}) \times \frac{N_m(k)}{\binom{m}{k}}$$

$$P_d(j) = P(\text{failures of } j \text{ storage devices}) \times \frac{N_d(j)}{\binom{mn}{j}}$$

$$P_{m,d}(k,j) = P(\text{failures of } k \text{ servers}) \times P(\text{failures of } j \text{ storage devices}) \times \frac{N_{m,d}(k,j)}{\binom{m}{k} \times \binom{mn}{j}}$$

The possibility of data loss of embedded storage systems

$$P(\text{failures of j storage devices}) = \frac{{R_d}^j e^{-R_d}}{j!}$$

$$P_{es} = \sum_{k=3}^{m'} P'_m(k) + \sum_{j=3}^{m'} P'_d(j) + \sum_{k=2}^{m'}\sum_{j=1}^{m'} P'_{m,d}(k,j) + \sum_{j=2}^{m'} P'_{m,d}(1,j)$$

where

$$P'_m(k) = \frac{{R'_m}^k e^{-R'_m}}{k!} \times \frac{N'_m(k)}{\binom{m'}{k}}$$

$$P'_d(j) = \frac{{R'_d}^j e^{-R'_d}}{j!} \times \frac{N'_d(j)}{\binom{m'}{j}}$$

$$P'_{m,d}(k,j) = \frac{{R'_m}^k e^{-R'_m}}{k!} \times \frac{{R'_d}^j e^{-R'_d}}{j!} \times \frac{N'_{m,d}(k,j)}{\binom{m'}{k} \times \binom{m'}{j}}$$