Tintenfisch: File System Namespace Schemas and Generators

Michael A. Sevilla, Reza Nasirigerdeh, Carlos Maltzahn, Jeff LeFevre, Noah Watkins, Peter Alvaro, Margaret Lawson*, Jay Lofstead*, Jim Pivarski^

UC Santa Cruz, *Sandia National Laboratories, ^Princeton University







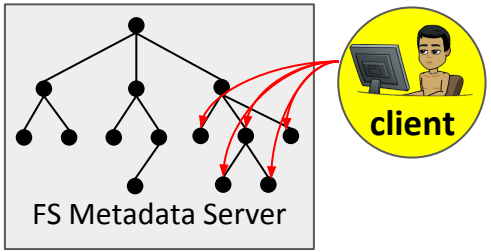




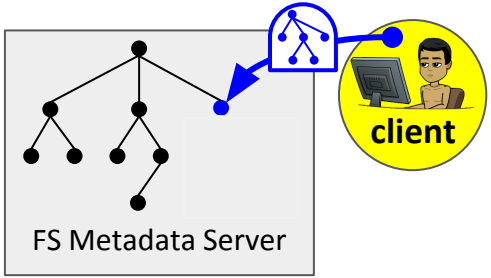
Sandia National Laboratories

Overview: Transfer/Materialization of FS Lists

Problem: RPCs $\forall_{ops}$ in **distributed file systems**



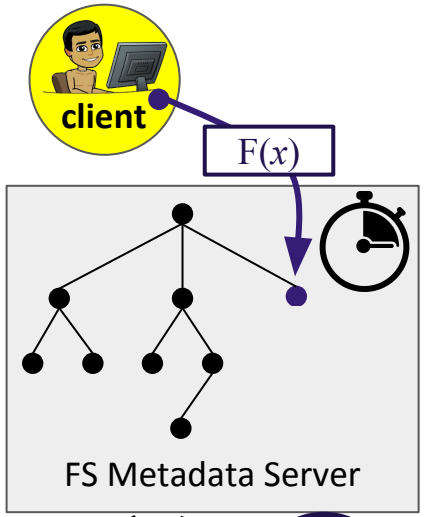
Related Work: Decouple



… but then **reads** must transferred & materialized

Tintenfisch:

1. High Perf. Computing
2. High Energy Physics
3. Large Simulations



- ✓ faster transfer
- ✓ faster modification
- ✓ faster generation

Schemas: Namespaces are LARGE but predictable (bounded/balanced)

Generators:

$F(x) = 2 \times n$

```
for(i=0;i<len;i++){
```

1 ✓ 3 ✓ 2 ✓

Outline

Tintenfisch: File System Namespace Schemas and Generators

Michael A. Sevilla
HotStorage '18
July 9, 2018

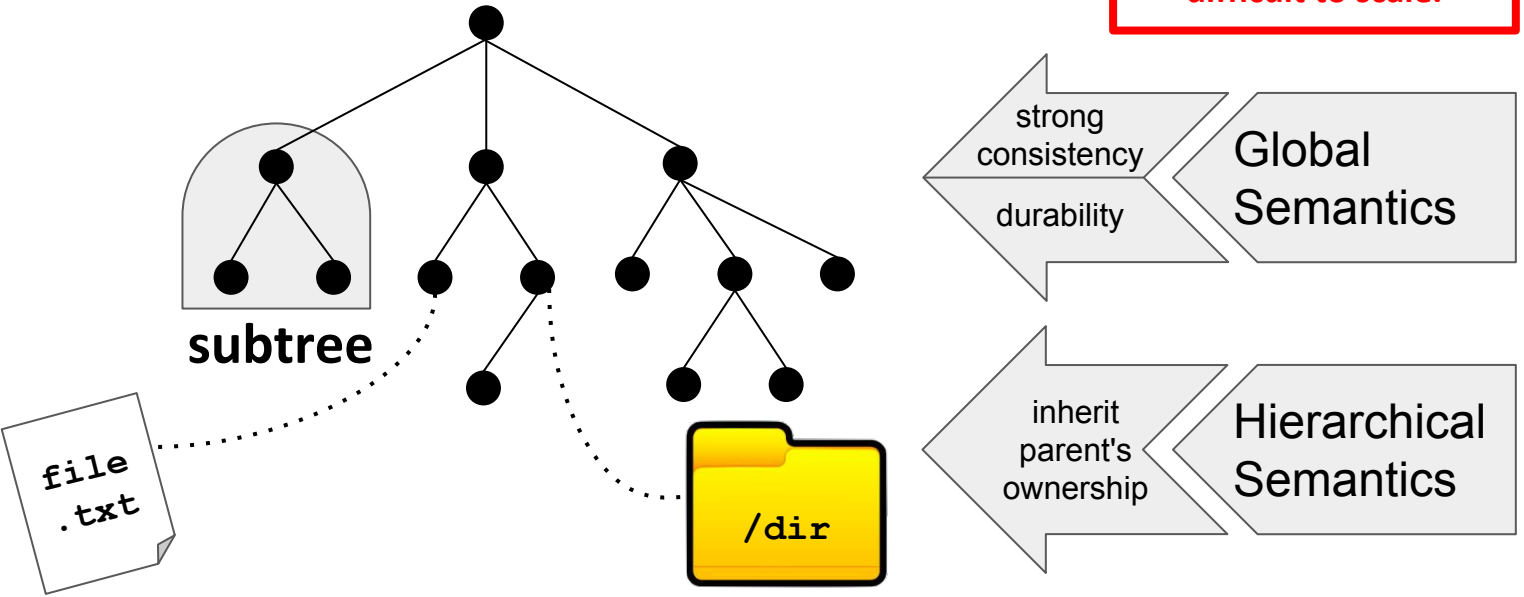
Introduction ←
Related Work
Motivating Examples
High Perf. Computing
High Energy Physics
Large Scale Simulation
Namespace Schemas
Namespace Generators
Conclusion



Primer on file system namespaces

Names ↦ Data; Hierarchical Structure

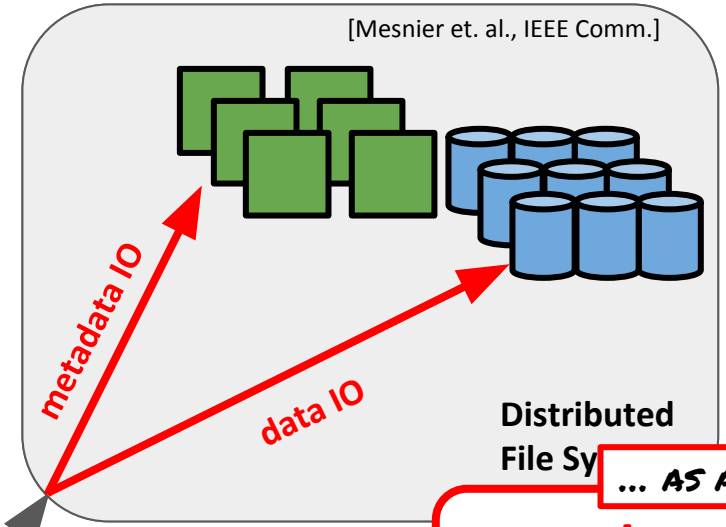
Problem! POSIX IO file system metadata access semantics are difficult to scale.



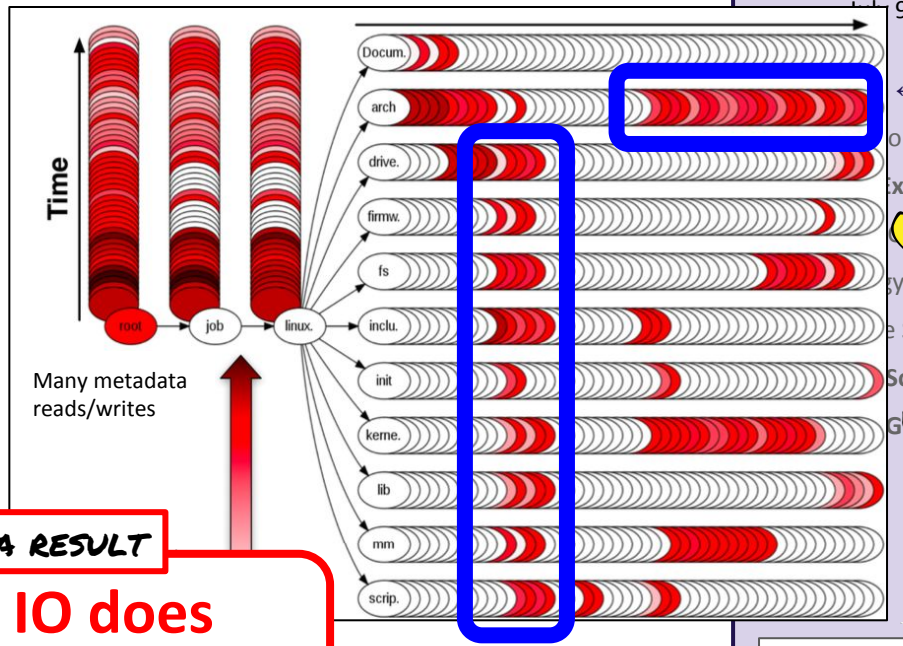
Introduction ←
Related Work
Motivating Examples
High Perf. Computing
High Energy Physics
Large Scale Simulation
Namespace Schemas
Namespace Generators
Conclusion

Primer on FS metadata

[Mesnier et. al., IEEE Comm.]



[Sevilla et. al., SC'15]



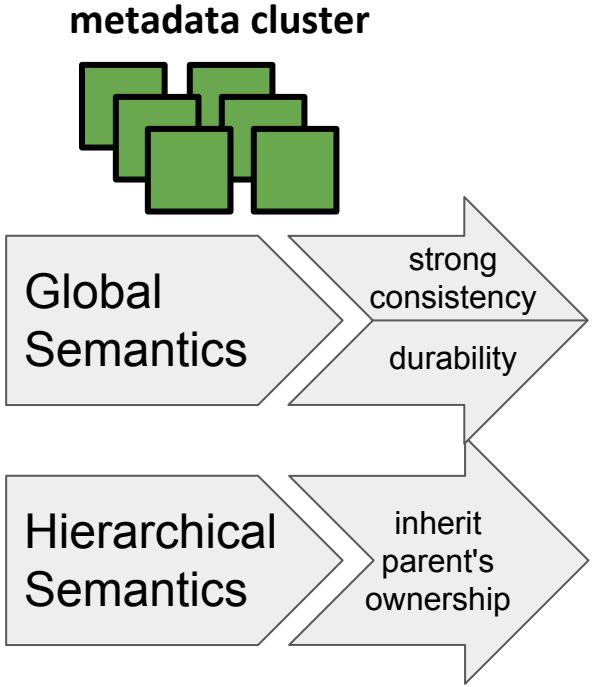
... AS A RESULT

metadata IO does not scale like data IO

File system metadata access patterns:

→ Small and frequent requests

→ Target same resource



- lock management
- relaxing consistency
- caching inodes
- journal formats
- journal safety
- caching paths
- metadata distribution
- load balancing

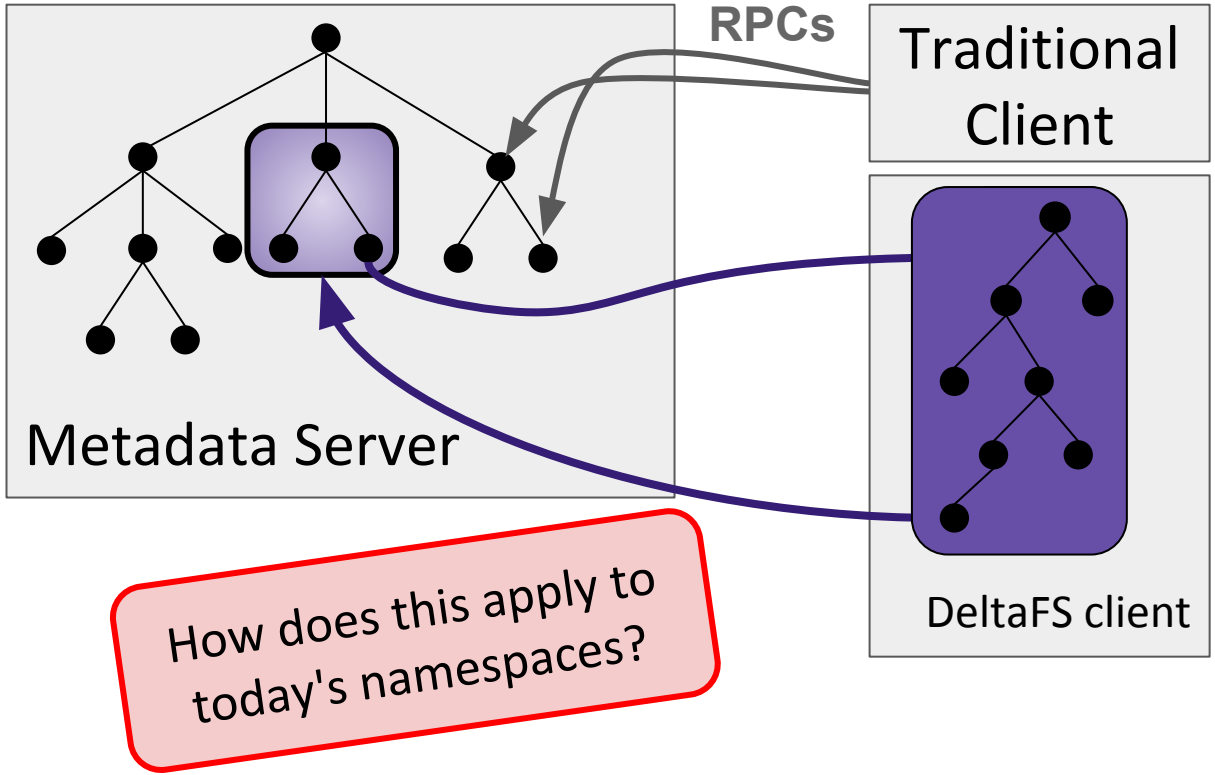
Michael A. Sevilla

HotStorage '18
July 9, 2018

Introduction
Related Work ←
Motivating Examples
High Perf. Computing
High Energy Physics
Large Scale Simulation
Namespace Schemas
Namespace Generators
Conclusion

Related Work: Decoupled Namespaces

[Zheng et al., PDSW'14; Zheng et al., PDSW'15]



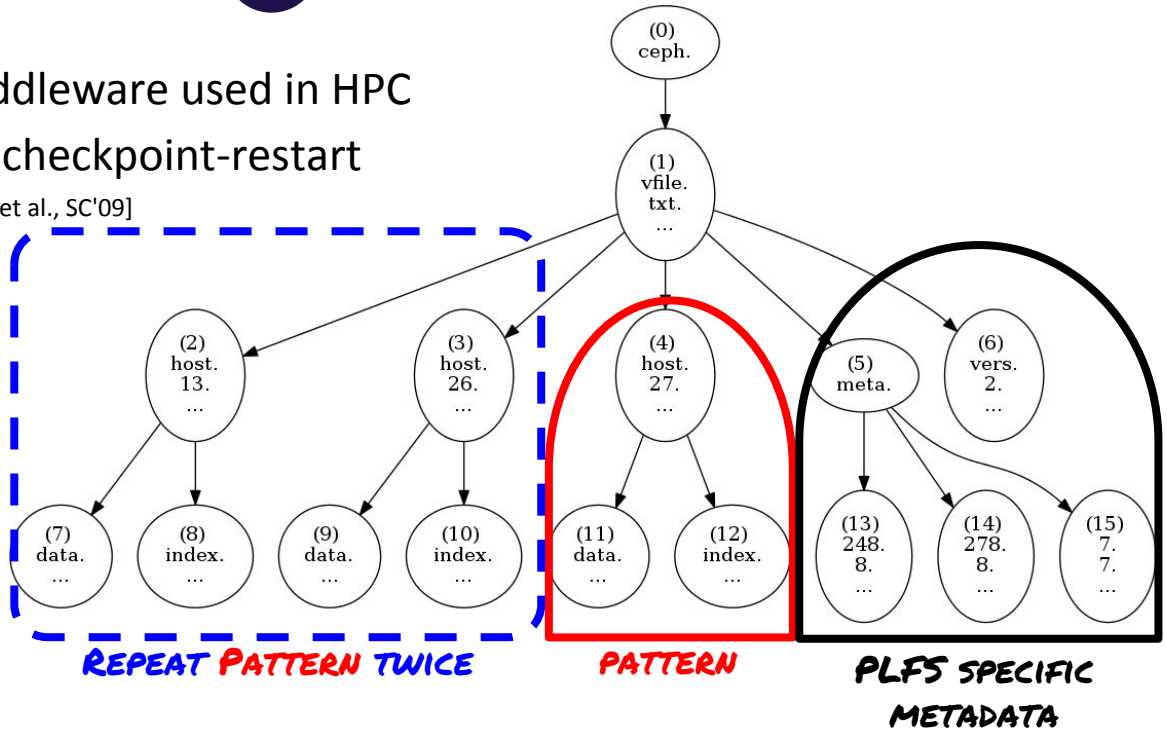
Introduction
Related Work ←
Motivating Examples
High Perf. Computing
High Energy Physics
Large Scale Simulation
Namespace Schemas
Namespace Generators
Conclusion

Example 1: PLFS Namespace

Middleware used in HPC

for checkpoint-restart

[Bent et al., SC'09]

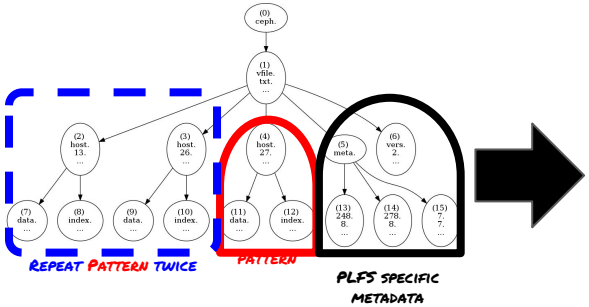


Example 1: PLFS Namespace

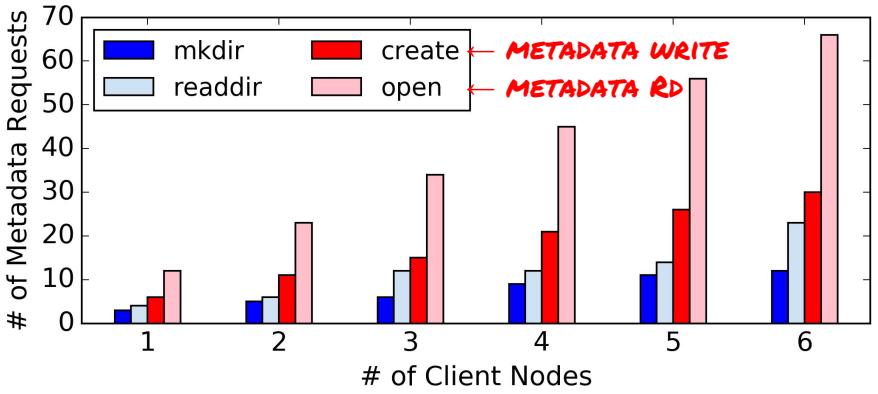
Middleware used in HPC

for checkpoint-restart

[Bent et al., SC'09]



- scales with # of clients

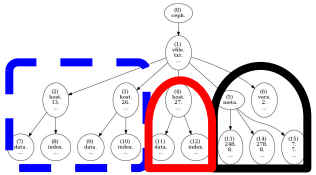


- RPCs probably not an option; decoupled ✓writes x reads (RD)

Introduction
Related Work
Motivating Examples
High Perf. Computing ←
High Energy Physics
Large Scale Simulation
Namespace Schemas
Namespace Generators
Conclusion

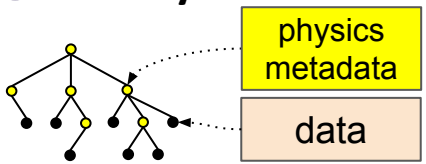
Summary of File System Namespaces

1 High Perf. Computing



- scales with # of clients

2 High Energy Physics

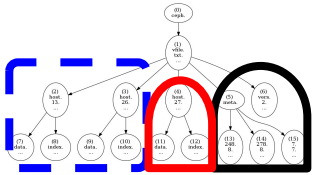


- ~ billions of files

Introduction
Related Work
Motivating Examples
High Perf. Computing
High Energy Physics ←
Large Scale Simulation
Namespace Schemas
Namespace Generators
Conclusion

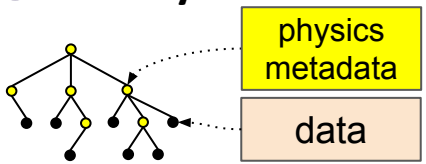
Summary of File System Namespaces

1 High Perf. Computing



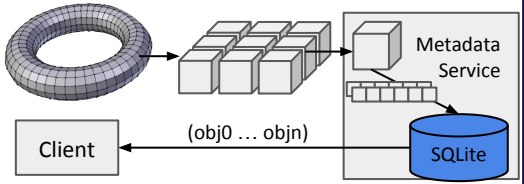
- scales with # of clients

2 High Energy Physics



- ~ billions of files

3 Large Scale Simulations

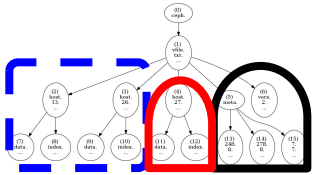


- ~ trillions of objects

Introduction
Related Work
Motivating Examples
High Perf. Computing
High Energy Physics
Large Scale Simulation ←
Namespace Schemas
Namespace Generators
Conclusion

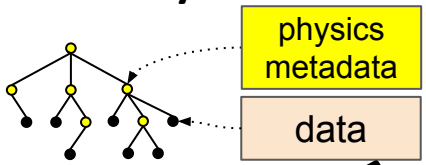
Summary of File System Namespaces

1 High Perf. Computing



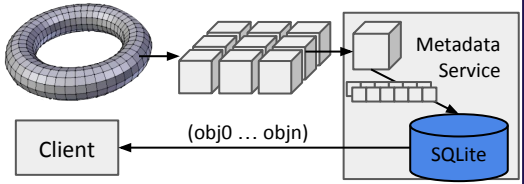
- scales with # of clients

2 High Energy Physics



- ~ billions of files

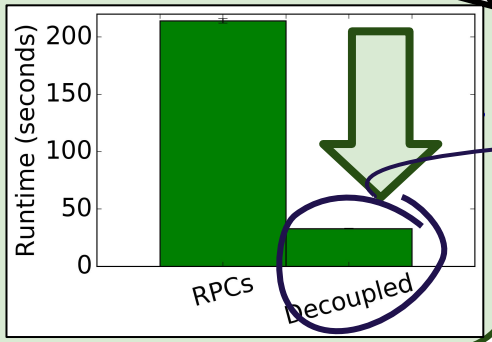
3 Large Scale Simulations

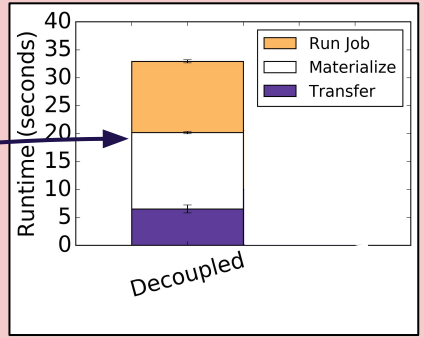


- ~ trillions of objects

example

Decoupling **definitely** improves performance



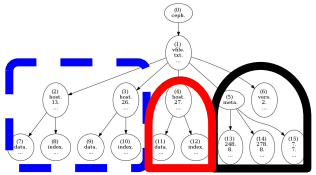


… but then **reads** must transferred & materialized

SANTA CRUZ
CENTER FOR RESEARCH IN OPEN SOURCE SOFTWARE

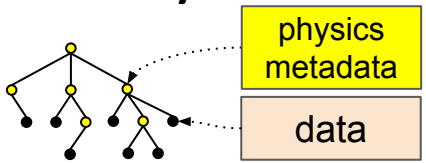
Summary of File System Namespaces

1 High Perf. Computing



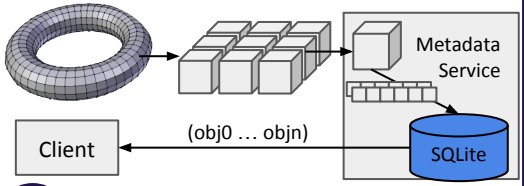
scales with # of clients

2 High Energy Physics



~ billions of files

3 Large Scale Simulations

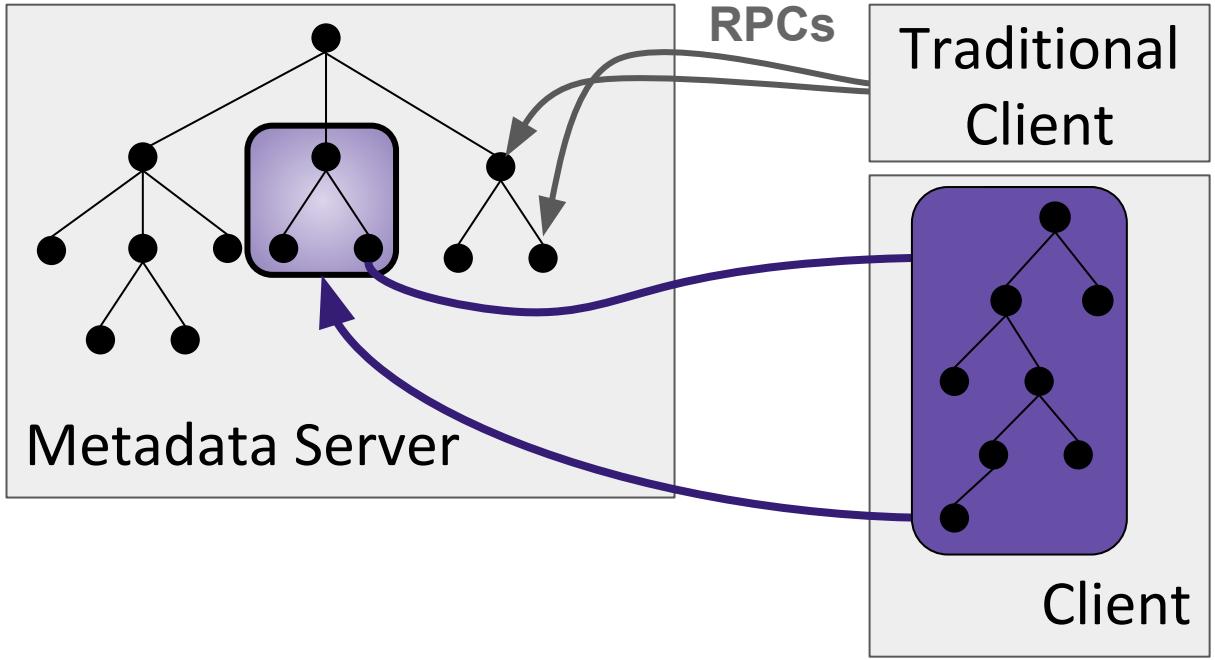


~ trillions of objects

Schemas: Namespaces are LARGE but predictable (bounded/balanced)

Integrate w/ decoupled namespaces

[Zheng et al., PDSW'15]



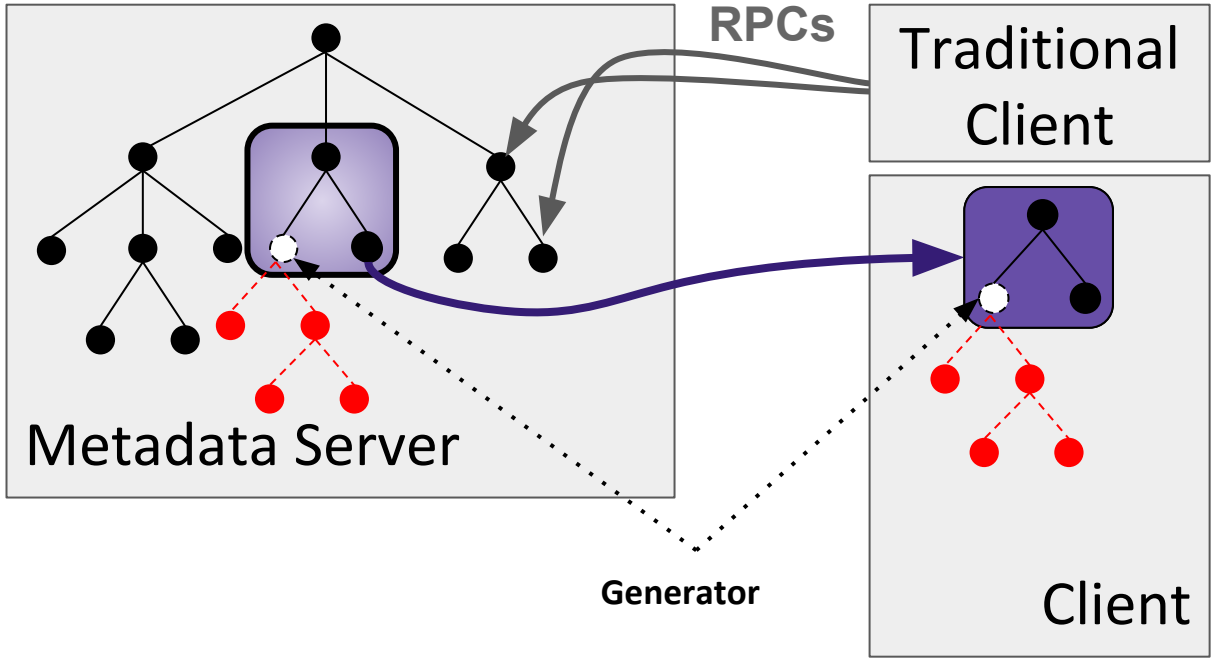
"Decoupled Namespace" policy implemented on ceph in Cudele

[Sevilla et. al., IPDPS'18]

Michael A. Sevilla
HotStorage '18
July 9, 2018

Introduction
Related Work
Motivating Examples
High Perf. Computing
High Energy Physics
Large Scale Simulation
Namespace Schemas ←
Namespace Generators
Conclusion

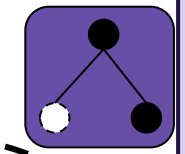
Tintenfisch builds on decoupled namespaces



"Decoupled Namespace" policy implemented on ceph in Cudele

[Sevilla et. al., IPDPS'18]

Namespace generators: compact metadata



Formula Generator

$$\text{Files}(n) = 2 \times n$$
$$\text{Dirs}(m) = m$$

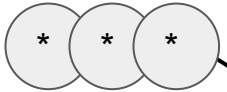
1 ✓ **High Perf. Computing**

Code Generator

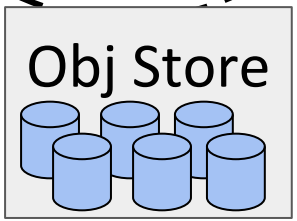
```
import bbox
if(t>30){
  obj=bbox.split(4)
```

3 ✓ **Fusion Simulation**

Pointer Generator



2 ✓ **High Energy Physics**



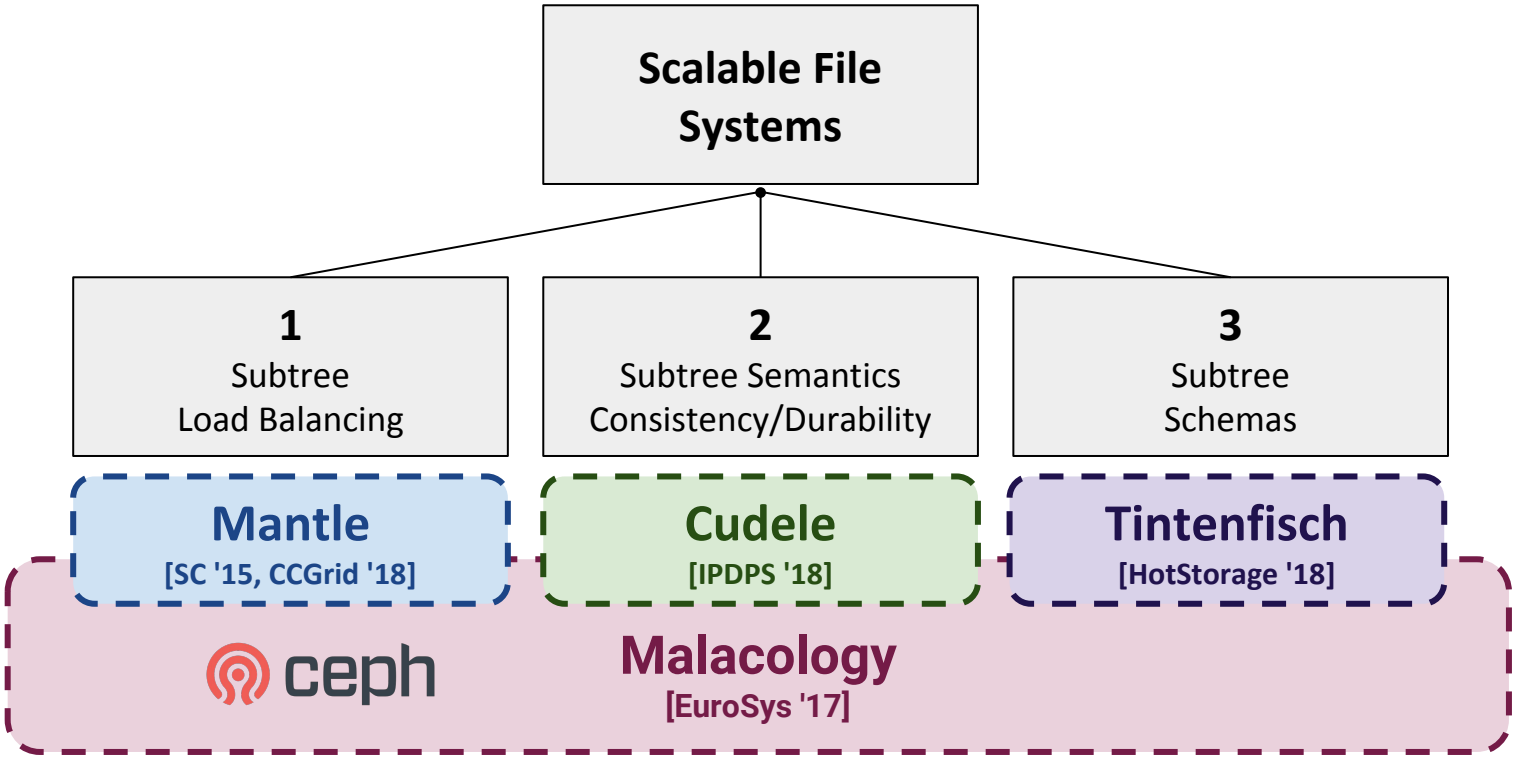
- ✓ faster transfer (metadata compaction)
- ✓ faster modification (change generator)
- ✓ faster generation (no more list/filter)

Michael A. Sevilla

HotStorage '18
July 9, 2018

Introduction
Related Work
Motivating Examples
High Perf. Computing
High Energy Physics
Large Scale Simulation
Namespace Schemas
Namespace Generators ←
Conclusion

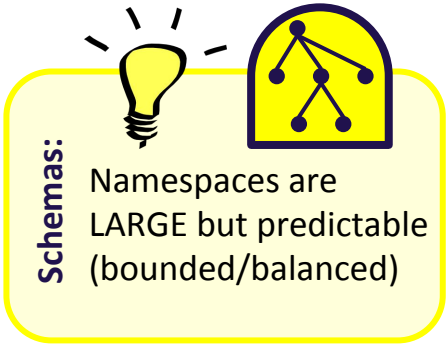
Prior work with programmable storage

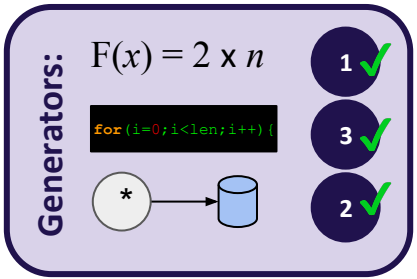


Introduction ←
Related Work
Motivating Examples
High Perf. Computing
High Energy Physics
Large Scale Simulation
Namespace Schemas
Namespace Generators
Conclusion

Conclusion: transferring and materializing reads

Related work: decoupled namespace overheads with large namespaces





- ✓ faster transfer
- ✓ faster modification
- ✓ faster generation

UC SANTA CRUZ CROSS CENTER FOR RESEARCH IN OPEN SOURCE SOFTWARE

Future Work

- proper sandboxing for security/correctness
- more complex file system metadata
  - permissions (workflows)
  - size of the file (Hadoop)
  - timestamps and dates (GC)
- storage system agnostic metadata generation

Thanks!

Tintenfisch: File System Namespace Schemas and Generators

Michael A. Sevilla, Reza Nasirigerdeh, Carlos Maltzahn, Jeff LeFevre, Noah Watkins, Peter Alvaro, Margaret Lawson*, Jay Lofstead*, Jim Pivarski^

UC Santa Cruz, *Sandia National Laboratories, ^Princeton University

More information:
- programmability.us








UC SANTA CRUZ

Sandia National Laboratories