SNIA DEVELOPER CONFERENCE

SDC 24

BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

Implementing Selective Write-Grouping

in SDS for Enhanced Energy Savings

Presented by
Piotr Modrzyk & David Gerstein

Speakers

Piotr Modrzyk

Principal Architect at Leil Storage
and SaunaFS
X-googler and Creator of LizardFS

David Gerstein

Founder & CTO at Leil Storage
and SaunaFS

Outline

- **Introduction**
- **Energy Efficiency in Large-Scale Storage Deployments**
  - Usual data access patterns
  - Common issues related to energy consumption
- **Existing Solutions**
  - Dynamic Power Management, Workload Skew, Cache, PDC...
- **Solution**
  - Selective Write-Grouping
- **Implementation Details in SaunaFS Distributed File System**
- **Results**

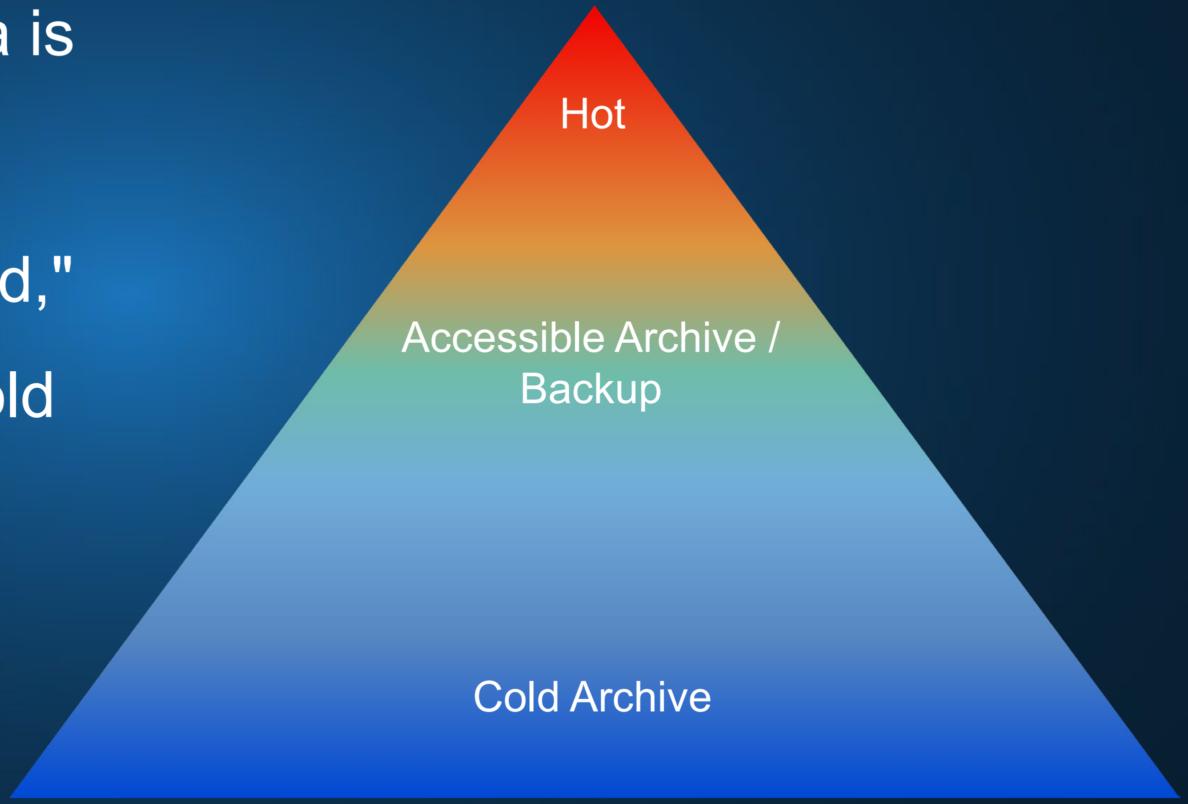
Introduction

- The extensive distribution of data across servers and drives presents significant challenges in terms of energy consumption.
- With many servers and drives running simultaneously, the overall energy demand becomes substantial.
- This leads to considerable environmental impact and increased operational costs for data centers.
- In large-scale storage systems, most of the energy consumption originates from the drives.

Energy Efficiency in Large-scale Storage Deployments

Usual Data Access Patterns

- In large-scale storage systems, not all data is accessed frequently.
- A large portion, around 60%[1], remains "cold," even if it's not designated as nearline or cold storage.
- For cold data, HDDs are the most cost-effective and energy-efficient option.



[1] Active Archive and the State of the Industry 2020

Energy Inefficiencies in RAID and EC Storage Systems

- Both Scale-up (RAID-based) and Scale-out (EC-based) storage solutions typically distribute data across all available drives.
- Servers and HDDs are usually kept running, even though the hardware is capable of being powered down.
- As a result, many drives remain active, even when a large portion of the data is "cold," causing significant energy inefficiency.
- This high energy consumption has both financial and environmental consequences.

Existing Solutions

Existing Energy-Saving Techniques

- Dynamic Power Management (DPM)
- Workload Skew (WS)
- DPM + WS
  - MAID (DPM + some WS)
  - Popular Data Concentration (WS + some DPM)
- RAID configurations or Erasure Coding vs replication
- Data Deduplication and Compression

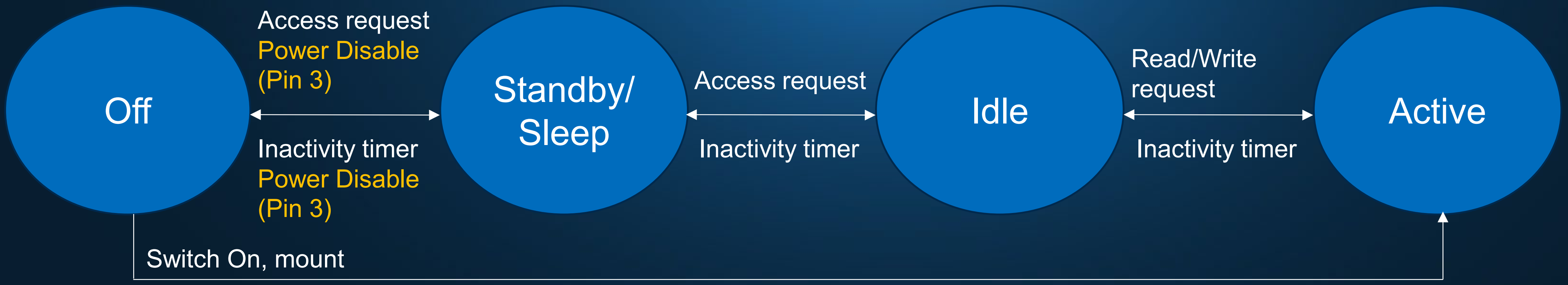
HDD Dynamic Power Management

Disk drives typically have multiple power states or levels, including:

- **Active**: Fully powered and at maximum performance.
- **Idle**: Powered on but not currently being accessed. Power consumption is reduced but can quickly return to the Active.
- **Standby/Sleep**: Low-power state, with its platters spun down or its circuitry in a reduced power mode.
- **Off**: Completely powered down.

Previous Energy-Saving Techniques

- Various energy-saving techniques effectively utilize the different power states of disk drives.
- By optimizing these transitions, significant energy savings can be achieved.



Challenges with Energy-Saving Techniques

While Dynamic Power Management (DPM) and Workload Skew (WS) are conceptually different, both aim to minimize the number of active disks to reduce power consumption.

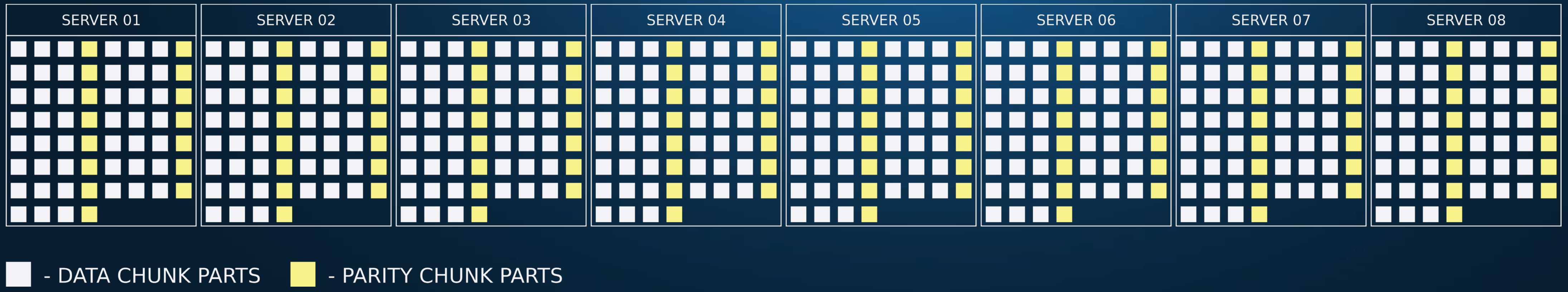
- **DPM** dynamically adjusts disk power states based on real-time usage, while **WS** strategically balances workloads to keep most disks underutilized.
- Despite being an industry standard, **DPM** struggles with current RAID and EC setups, where data is distributed across all available disks, reducing its energy-saving efficiency. **DPM** works best when disk usage is minimal.
- To fully leverage **DPM**, optimizing data distribution across storage is essential.

Solution

Selective Write-Grouping

Our Typical Deployments

- 8 Servers, each connected to a JBOD with 60 or 102 drives.
- Drives can be SMR, CMR, SSD or NVMe (or mixed).
- Most of them support the Power Disable feature through Pin 3.
- Erasure coding with 6 data parts (blue) and 2 parities (yellow), aka EC(6,2).



Our Typical Deployments

A typical drive specification:

- Serie: Ultrastar DC HC680.
- Model: WSH722880ALN604
- 28TB, 7200 RPM, SATA 3.3.
- Base (SE) configuration
- Host-Managed SMR with 256 MiB zone size.
- P3 "Power Disable" supported.

Our Typical Deployments

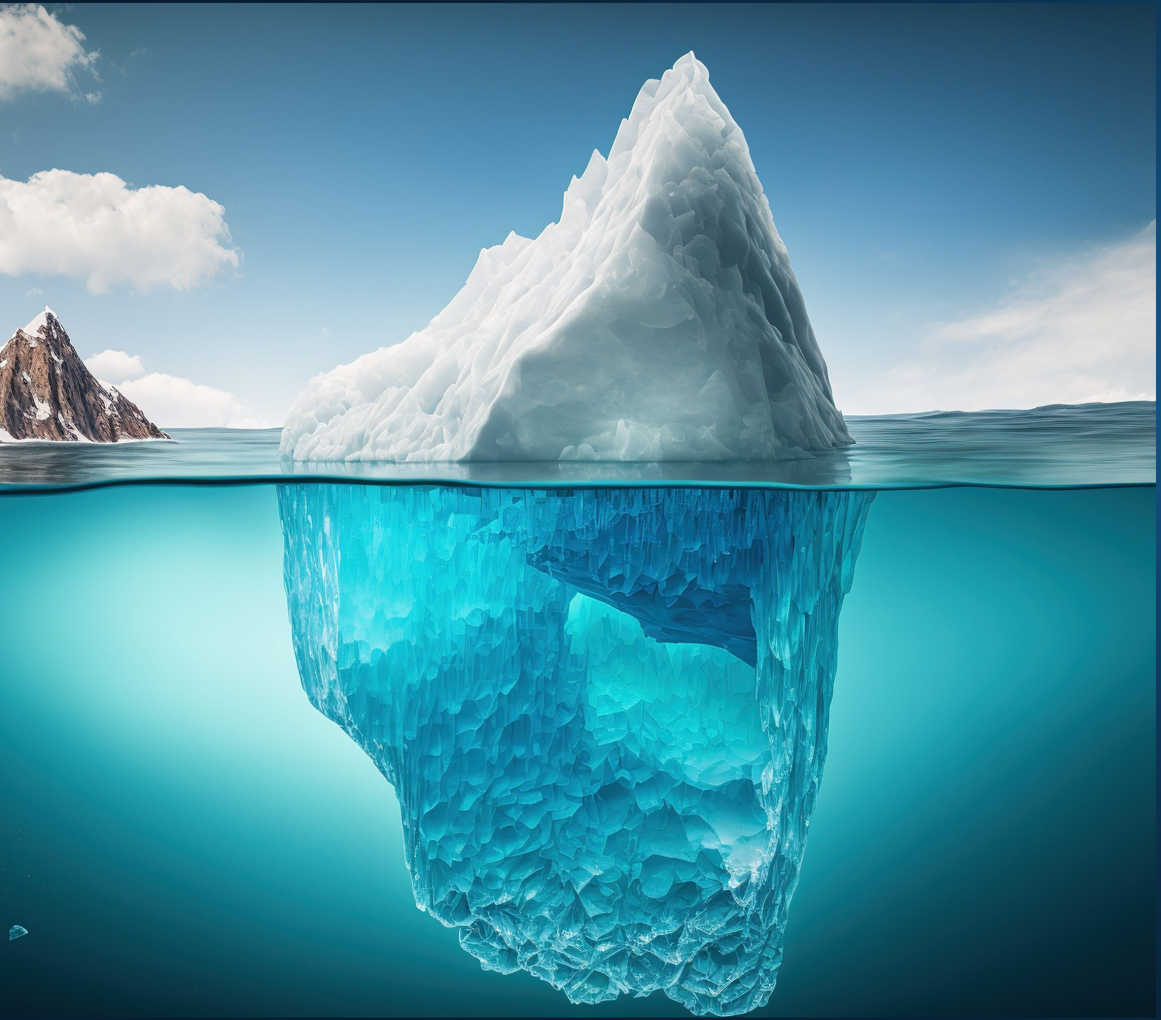
Power condition	Power (W)	Description
Active (at max workload)	9.4	Ready to perform IO immediately.
Idle_0	5.5	Ready but not doing IO, may power down selected electronics.
Idle_A	5.5	Ready but not doing IO, may power selected down electronics.
Idle_B	3.7	Spindle rotation at 7200 RPM with heads unloaded.
Idle_C/Standby_Y	3.2	Spindle rotation at Low RPM with heads unloaded.
Standby_Z	1.2	Actuator unloaded and spindle motor stopped.
Sleep	1.2	Actuator unloaded and spindle motor stopped. Only soft reset or hard reset can change the mode to Standby_Z.
Off	0.0	Drive is completely turned off.

Our Typical Deployments

Mode transition times

From	To	RPM	Typical (sec)
Idle_B	Active	7200	1.5
Idle_C/Standby_Y	Active	6300 -> 7200	4
Standby_Z	Active	0 -> 7200	15
Off	Active	0 -> 7200	15-30

ICE as Our Take on MAID (DPM* + WS**)



Infinite Cold Engine (ICE)

- Phase 1: HM-SMR Support **18% total energy savings**
- **Phase 2: Write-Grouping Conception Y2024 43%*****
- Phase 3: Smart Data Placement for WG Y2025 **50%*****
- Phase 4: AI-Driven Background Service Y2026 **70%*****

DPM* Dynamic Power Management

WS** Workload Skew

%*** Projected total energy savings

SDC 24

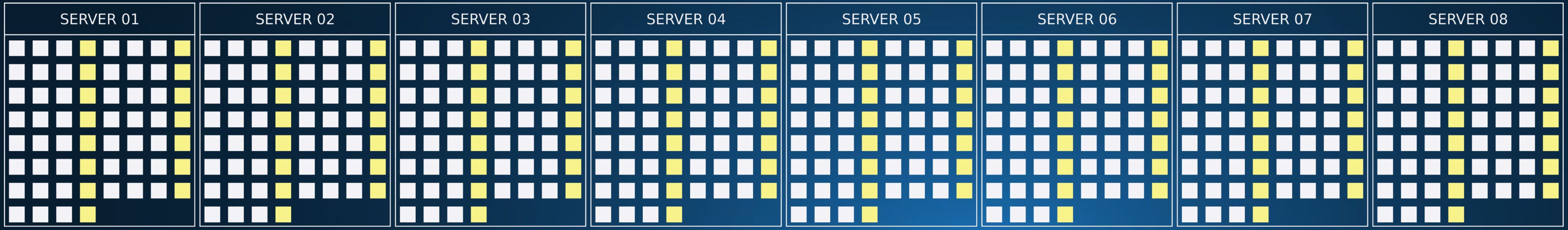
Phase 1: HM-SMR Support

Our first step was to add support for Host-Managed SMR drives, allowing:

- Higher density storage.
- Around a 10-20% more capacity with the same number of drives.
- Better data alignment.
- Chunks are now split into metadata and data parts.
- Better oriented to sequential writing.

SDC 2023: Bridging the Gap Between Host Managed SMR Drives and Software-Defined Storage

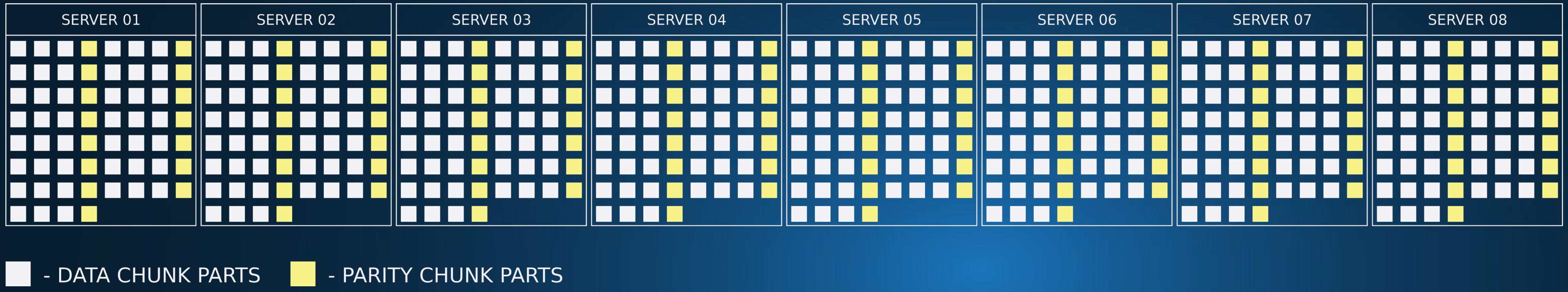
Phase 2: Selective Write-Grouping



- DATA CHUNK PARTS   - PARITY CHUNK PARTS

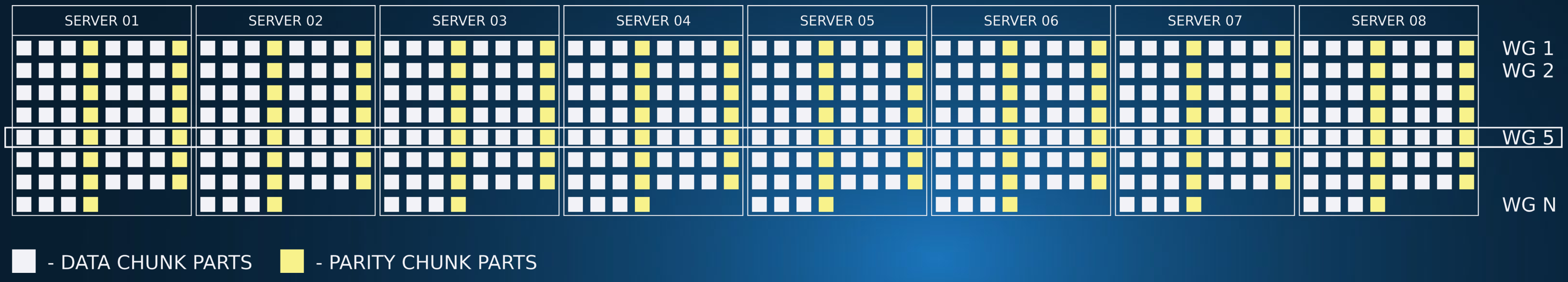
- As stated before, typical installations contains hundreds of HDDs.
  - Usually 8 * (60 or 102) = 480 – 816 drives.
- The original implementation distributes the data (Chunks) in a balanced way among all available disks.
- Disks with more available space are more probable to hold new data.

Phase 2: Selective Write-Grouping



- This balanced strategy makes difficult to keep the disks in a low power mode for long time periods.
- Causing frequent switches to Active mode in most of the disks.

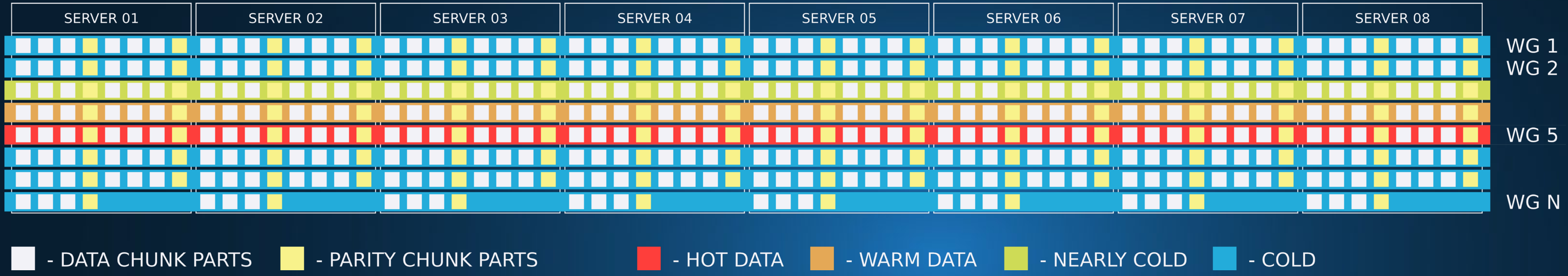
Phase 2: Selective Write-Grouping



Configurable Write Groups for Disk Distribution:

- Disks are organized into Write Groups, with one group designated as the Active Write Group for current write operations and with each disk assigned to a group.
- New data is typically written to the Active group, with rare exceptions like modifying or deleting old data, which is uncommon in archival storage.

Phase 2: Selective Write-Grouping



- The Active group (red) switches when space runs low, allowing most drives to stay in energy-saving modes or be powered down.
- Parity drives are only used for reconstruction, and full groups enter maximum power-saving mode, reducing energy use by 25%.

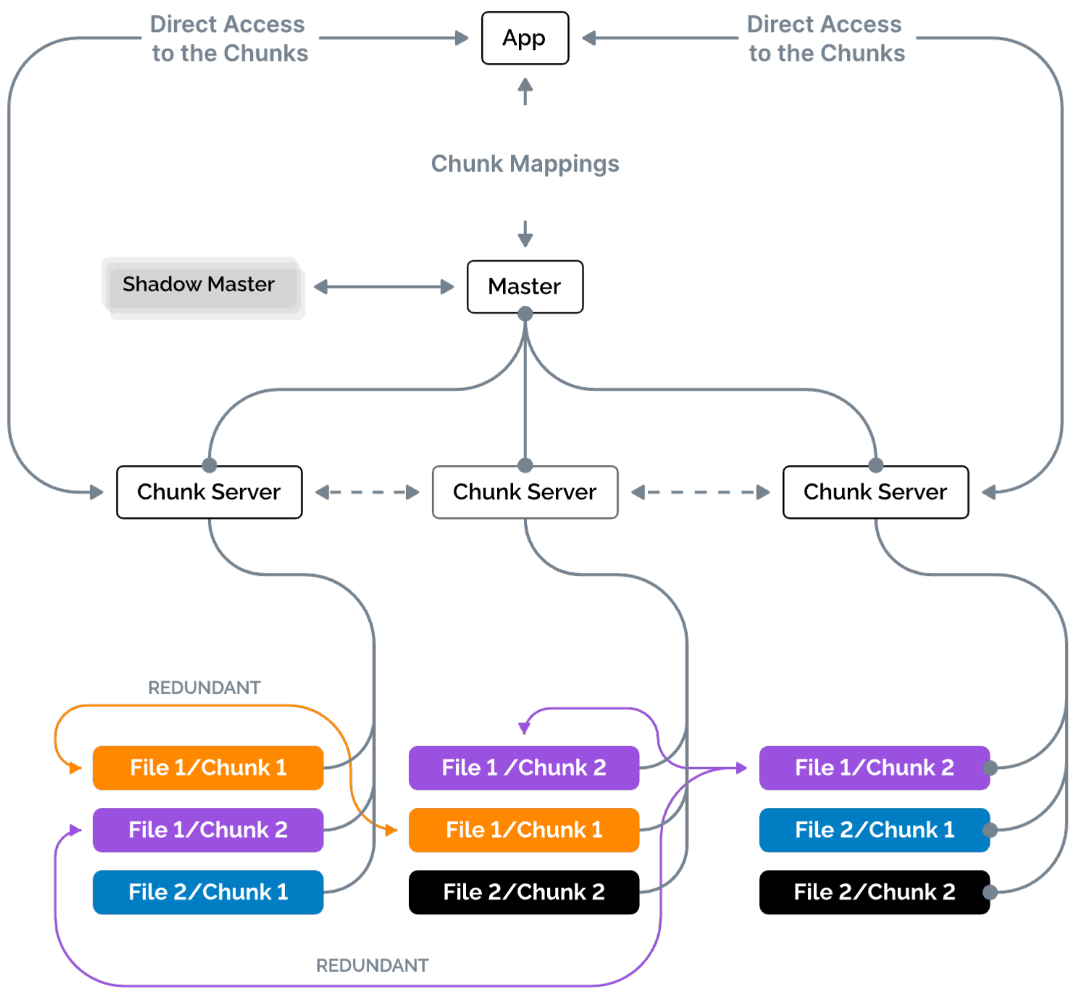
Implementation Details in SaunaFS Distributed File System

Brief Introduction to SaunaFS

SaunaFS, a C++-based DFS inspired by the Google File System, includes:

- Metadata Servers (master, shadows and metaloggers).
- Data Servers (chunkservers).
- Clients (native Linux/Windows, NFS).

Files are divided into 64 MiB chunks, further split into 64 KiB minimal blocks.



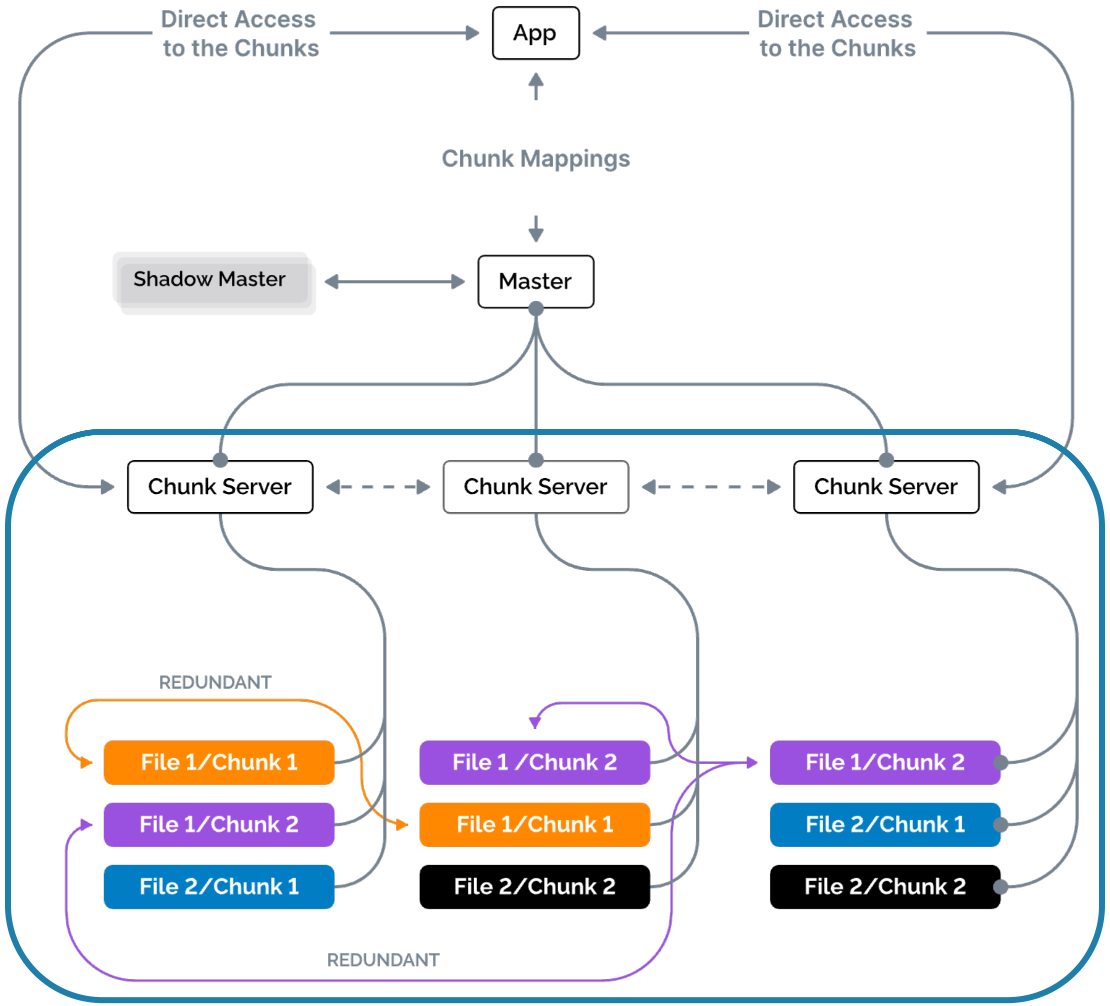
See https://en.wikipedia.org/wiki/Google_File_System

 |

SaunaFS Chunkserver

Chunkserver Responsibilities

- Stores data as Chunks and updates MDS servers.
- Pluggable architecture for easy extensions.
- Features include integrity checks, garbage collection, and Chunk replication or reconstruction.



See https://en.wikipedia.org/wiki/Google_File_System

Chunkserver Modifications

Disk configuration file previously had a plain text format, with each line defining a disk. Metadata and data parts of Chunks are stored separately:

- Metadata can be stored in NVMe (usually 4 KiB).
- Data is stored in HDDs (up to 64 MiB).

```
zonefs:/mnt/saunafs/meta/scsi-SATA_WDC_WSH722626AL_2TG3JRXF | /mnt/saunafs/data/scsi-SATA_WDC_WSH722626AL_2TG3JRXF
zonefs:/mnt/saunafs/meta/scsi-SATA_WDC_WSH722626AL_2TG3MJ6F | /mnt/saunafs/data/scsi-SATA_WDC_WSH722626AL_2TG3MJ6F
zonefs:/mnt/saunafs/meta/scsi-SATA_WDC_WSH722626AL_2HG1LLVN | /mnt/saunafs/data/scsi-SATA_WDC_WSH722626AL_2HG1LLVN
zonefs:/mnt/saunafs/meta/scsi-SATA_WDC_WSH722626AL_2TG375DF | /mnt/saunafs/data/scsi-SATA_WDC_WSH722626AL_2TG375DF
zonefs:/mnt/saunafs/meta/scsi-SATA_WDC_WSH722626AL_2GG75VNE | /mnt/saunafs/data/scsi-SATA_WDC_WSH722626AL_2GG75VNE
zonefs:/mnt/saunafs/meta/scsi-SATA_WDC_WSH722626AL_2TG3JNAF | /mnt/saunafs/data/scsi-SATA_WDC_WSH722626AL_2TG3JNAF
zonefs:/mnt/saunafs/meta/scsi-SATA_WDC_WSH722626AL_2HG1WN8N | /mnt/saunafs/data/scsi-SATA_WDC_WSH722626AL_2HG1WN8N
zonefs:/mnt/saunafs/meta/scsi-SATA_WDC_WSH722626AL_2GGMH5DT | /mnt/saunafs/data/scsi-SATA_WDC_WSH722626AL_2GGMH5DT
```


Chunkserver Modifications

SaunaFS

Disks configuration file:

- Due to the increased complexity, the disk configuration now uses YAML format.
- The structure separates data_disks and parity_disks, allowing parity disks (25%) to be powered down when not in use.
- Write Groups include a 'type' property, reserved for future development phases.


```yaml
version: 1.0
write_groups:
  - id: staging-area-1
    type: StagingArea
    data_disks:
      - disk: "/mnt/ramdisk/hdd_0_0"
      - disk: "/mnt/ramdisk/hdd_0_1"
      - disk: "/mnt/ramdisk/hdd_0_2"
    parity_disks:
      - disk: "/mnt/ramdisk/hdd_0_3"
  - id: staging-area-2
    type: StagingArea
    data_disks:
      - disk: "/mnt/ramdisk/hdd_0_4"
      - disk: "/mnt/ramdisk/hdd_0_5"
      - disk: "/mnt/ramdisk/hdd_0_6"
    parity_disks:
      - disk: "/mnt/ramdisk/hdd_0_7"
  - id: write-group-1
    type: WriteGroup
    data_disks:
      - disk: "/mnt/ramdisk/hdd_0_8"
      - disk: "/mnt/ramdisk/hdd_0_9"
    parity_disks:
      - disk: "/mnt/ramdisk/hdd_0_10"
  - id: write-group-2
    type: WriteGroup
    data_disks:
      - disk: "/mnt/ramdisk/hdd_0_11"
      - disk: "/mnt/ramdisk/hdd_0_12"
    parity_disks:
      - disk: "/mnt/ramdisk/hdd_0_13"
```


SDC 24

Chunkserver Modifications

Testing Framework:

- Now supports generating YAML configuration files.
- To simulate available space in a Write Group, on-demand NullBlk emulated devices can be created for testing.
- A new variable, **MIN_WRITE_GROUP_PERCENT_AVAIL**, has been added to the configuration file.

```
# No need for staging areas for this test
ICE_NumberOfStagingAreas=0

# Write groups configuration
ICE_NumberOfWriteGroups=4
ICE_WriteGroupWidth=3
ICE_ParityDisksPerWriteGroup=1
```

Definition in the test

```
version: 1.0
write_groups:
  - id: write-group-1
    type: WriteGroup
    data_disks:
      - disk: "zonefs:/mnt/ramdisk/metadata/sauna_nullb0 | /mnt/zoned/sauna_nullb0"
      - disk: "zonefs:/mnt/ramdisk/metadata/sauna_nullb1 | /mnt/zoned/sauna_nullb1"
    parity_disks:
      - disk: "zonefs:/mnt/ramdisk/metadata/sauna_nullb2 | /mnt/zoned/sauna_nullb2"
  - id: write-group-2
    type: WriteGroup
    data_disks:
      - disk: "zonefs:/mnt/ramdisk/metadata/sauna_nullb3 | /mnt/zoned/sauna_nullb3"
      - disk: "zonefs:/mnt/ramdisk/metadata/sauna_nullb4 | /mnt/zoned/sauna_nullb4"
    parity_disks:
      - disk: "zonefs:/mnt/ramdisk/metadata/sauna_nullb5 | /mnt/zoned/sauna_nullb5"
  - id: write-group-3
    type: WriteGroup
    data_disks:
      - disk: "zonefs:/mnt/ramdisk/metadata/sauna_nullb6 | /mnt/zoned/sauna_nullb6"
      - disk: "zonefs:/mnt/ramdisk/metadata/sauna_nullb7 | /mnt/zoned/sauna_nullb7"
    parity_disks:
      - disk: "zonefs:/mnt/ramdisk/metadata/sauna_nullb8 | /mnt/zoned/sauna_nullb8"
  - id: write-group-4
    type: WriteGroup
    data_disks:
      - disk: "zonefs:/mnt/ramdisk/metadata/sauna_nullb9 | /mnt/zoned/sauna_nullb9"
      - disk: "zonefs:/mnt/ramdisk/metadata/sauna_nullb10 | /mnt/zoned/sauna_nullb10"
    parity_disks:
      - disk: "zonefs:/mnt/ramdisk/metadata/sauna_nullb11 | /mnt/zoned/sauna_nullb11"
```

Generated configuration

Chunkserver Modifications

New specialized DiskManager:

- Now aware of Write Groups.
- Selects disks for new Chunks from the Active group, switching to another when space runs low.
- Provides group status updates for admin or monitoring tools.
- Manages power state transitions for disks.
- Uses Power Disable (P3) to power disks on or off after a defined period of inactivity.

Chunkserver Modifications

SaunaFS

New specialized DiskManager:

- For drives requiring Garbage Collection (GC), such as HM-SMR drives, the DiskManager ensures disks are selected from active Write Groups.
- For HM-SMR drives, GC involves defragmenting Chunks across multiple zones and recovering unused space by resetting the zones.

SDC 2023: Bridging the Gap Between Host Managed SMR Drives and Software-Defined Storage

SDC 24

Chunkserver Modifications

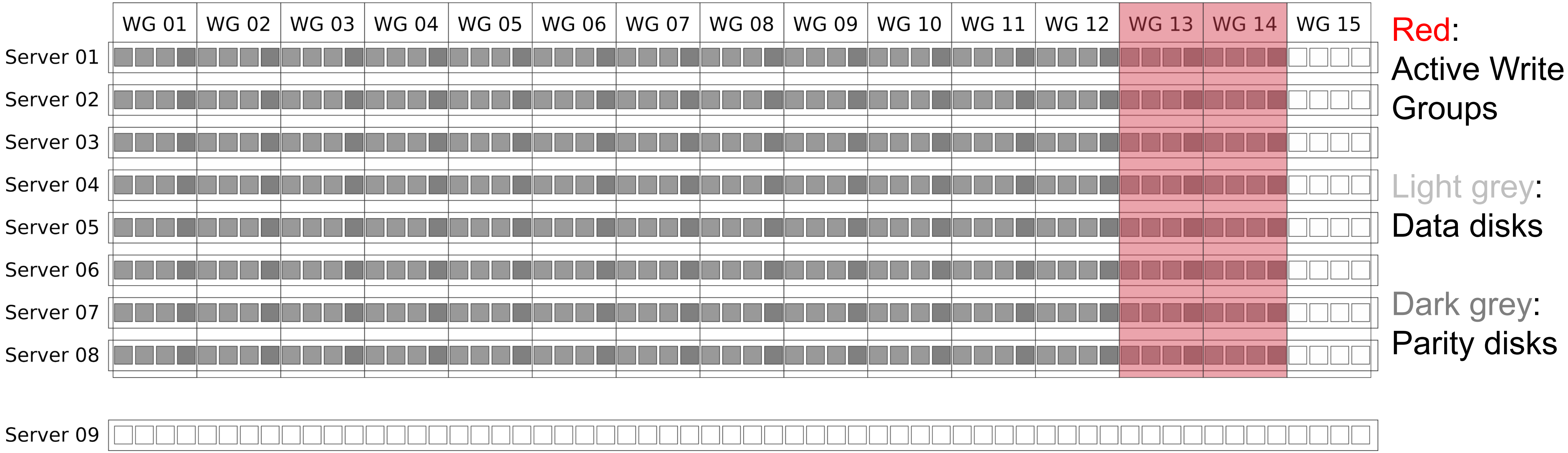
Rebalancing:

- The master server balances space usage across Chunkservers by replicating and removing Chunks.
- New Chunks are placed in the active Write Group, but rebalancing may wake up non-active disks if the original Chunk is on an inactive drive.

Chunk Servers			
'regular' hdd space			
chunks	used	total	% used
2252	55 GiB	189 TiB	0.03
2252	55 GiB	189 TiB	0.03
2254	55 GiB	189 TiB	0.03
2251	55 GiB	189 TiB	0.03
2245	51 GiB	166 TiB	0.03
2246	51 GiB	166 TiB	0.03
2245	51 GiB	166 TiB	0.03
2246	51 GiB	166 TiB	0.03

Rebalancing by Capacity Expansion

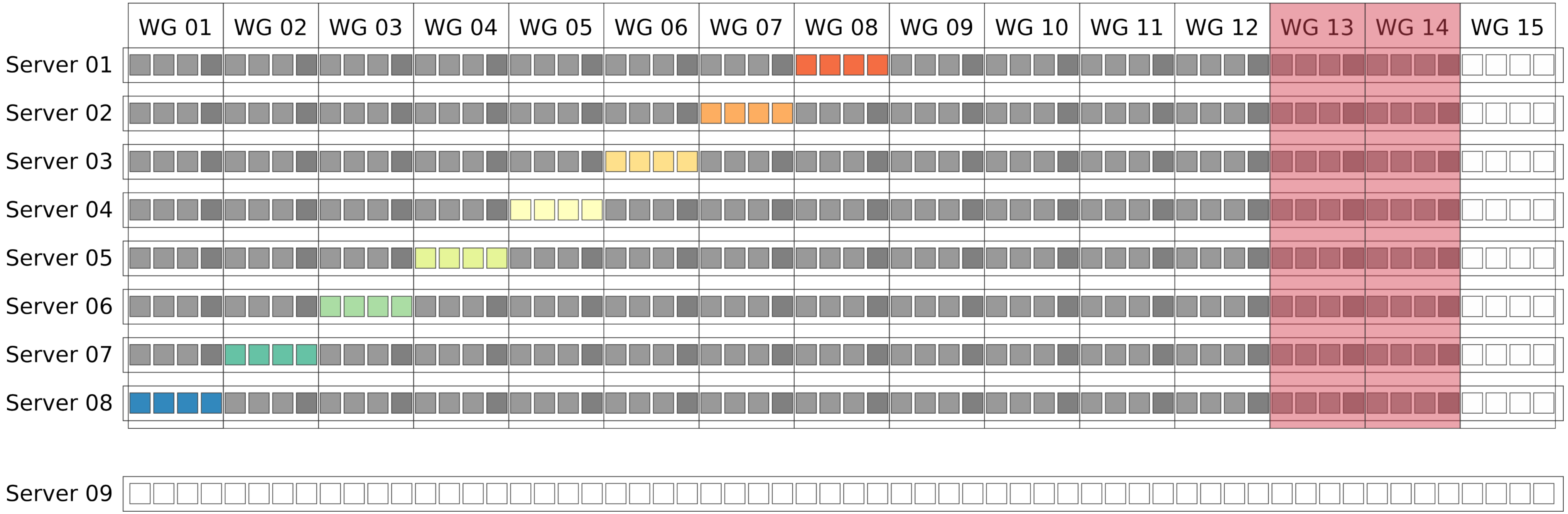
Adding a new server may trigger massive rebalancing, potentially disrupting EC(6,2) and data/parity drive assignments.



Rebalancing by Capacity Expansion

SaunaFS

To speed up rebalancing, we can safely copy or move rows from different Write Groups across multiple Chunkservers, visualized as a diagonal pattern.

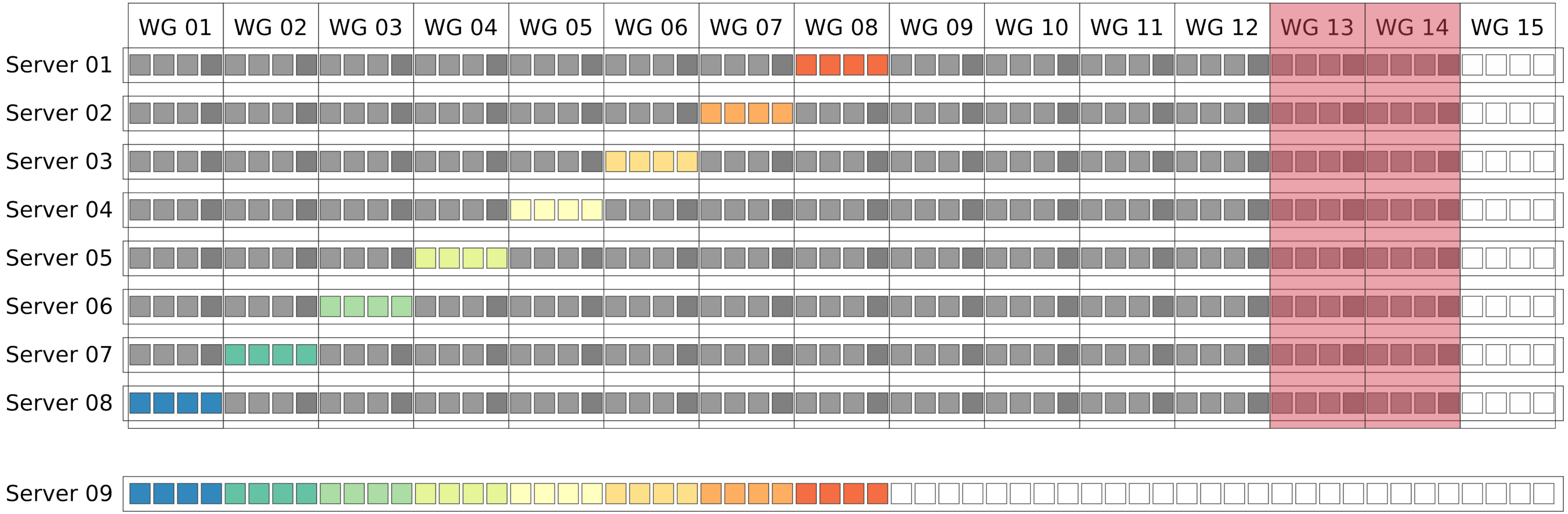


SDC 24

Rebalancing by Capacity Expansion

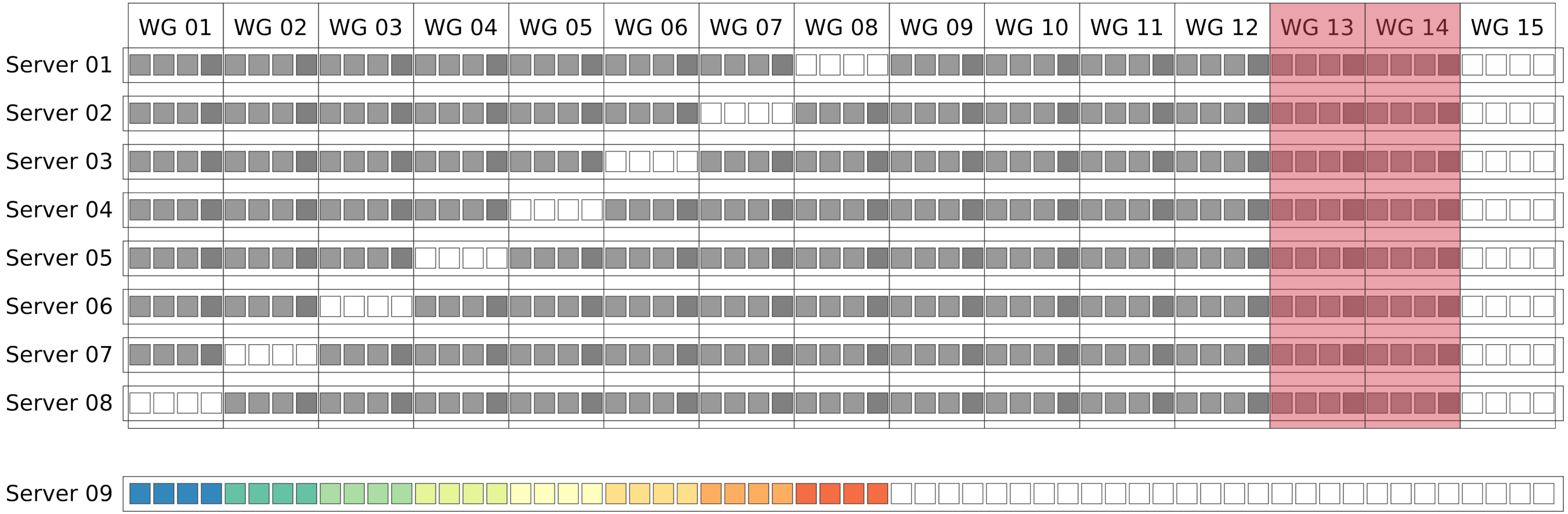
SaunaFS

This way, two pieces of the same EC(6,2) Chunk will NOT land in the same server.



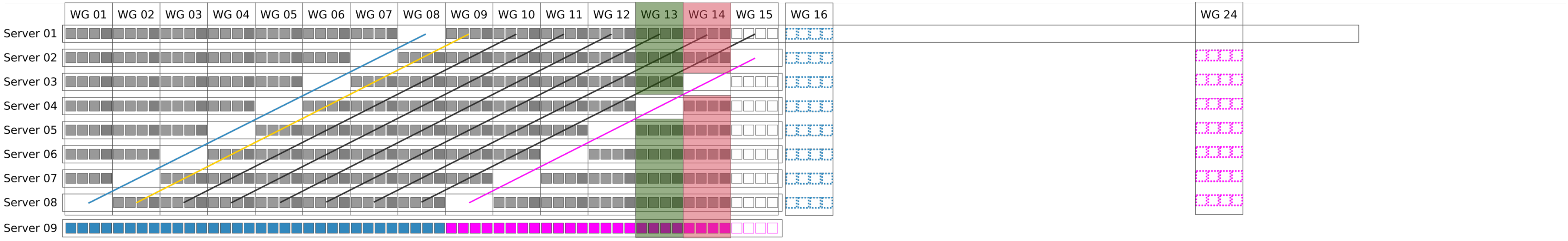
Rebalancing by Capacity Expansion

Empty drives can now become a new group.



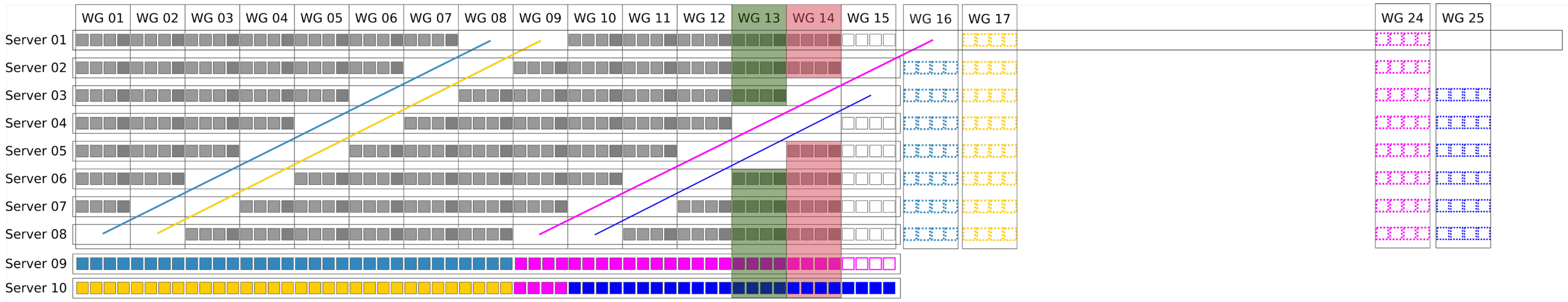
Rebalancing by Capacity Expansion

- A new incomplete diagonal (7 of 8, magenta) is added since server 09 has a capacity of only 60 drives in this example.
- Currently, WG 24 is the only unusable space (5.18%).



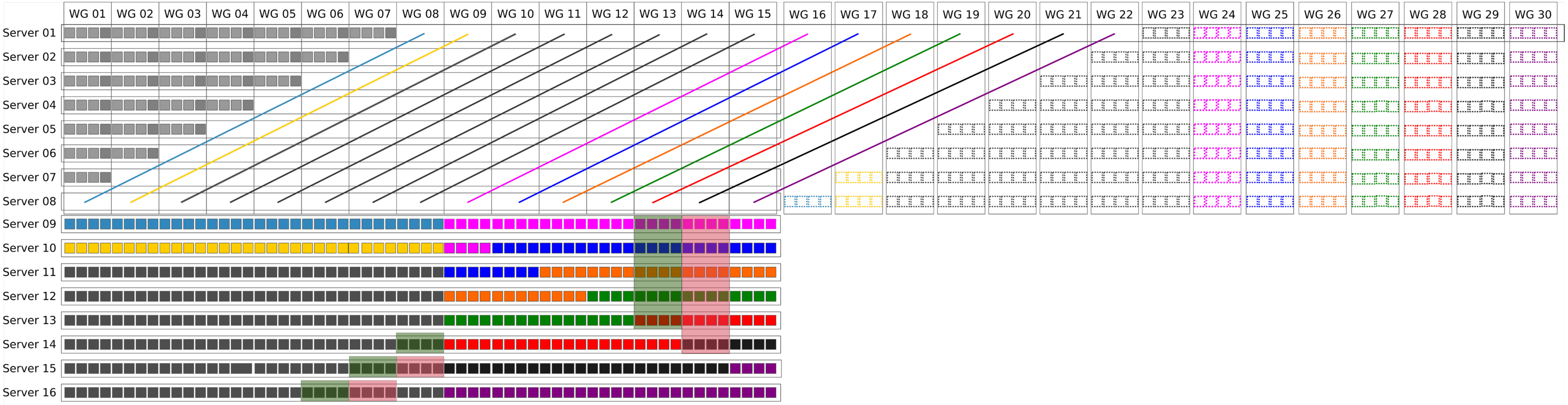
Rebalancing by Capacity Expansion

- The same process can be repeated for the second server.
- With empty disks, WG 24 can now be completed, and the unusable space shifts to WG 25 (4.00%).



Rebalancing by Capacity Expansion

- The diagram shows the full process for 8 new servers.
- Each part of the original Write Groups is placed on a different server, with no unusable space.



”Golden Gates” Approach for Data Scrubbing

During continuous data scrubbing, we can take advantage by performing:

- Garbage collection.
- Disk rotation, swapping inactive disks with those that were active for the past 6 months.

Results

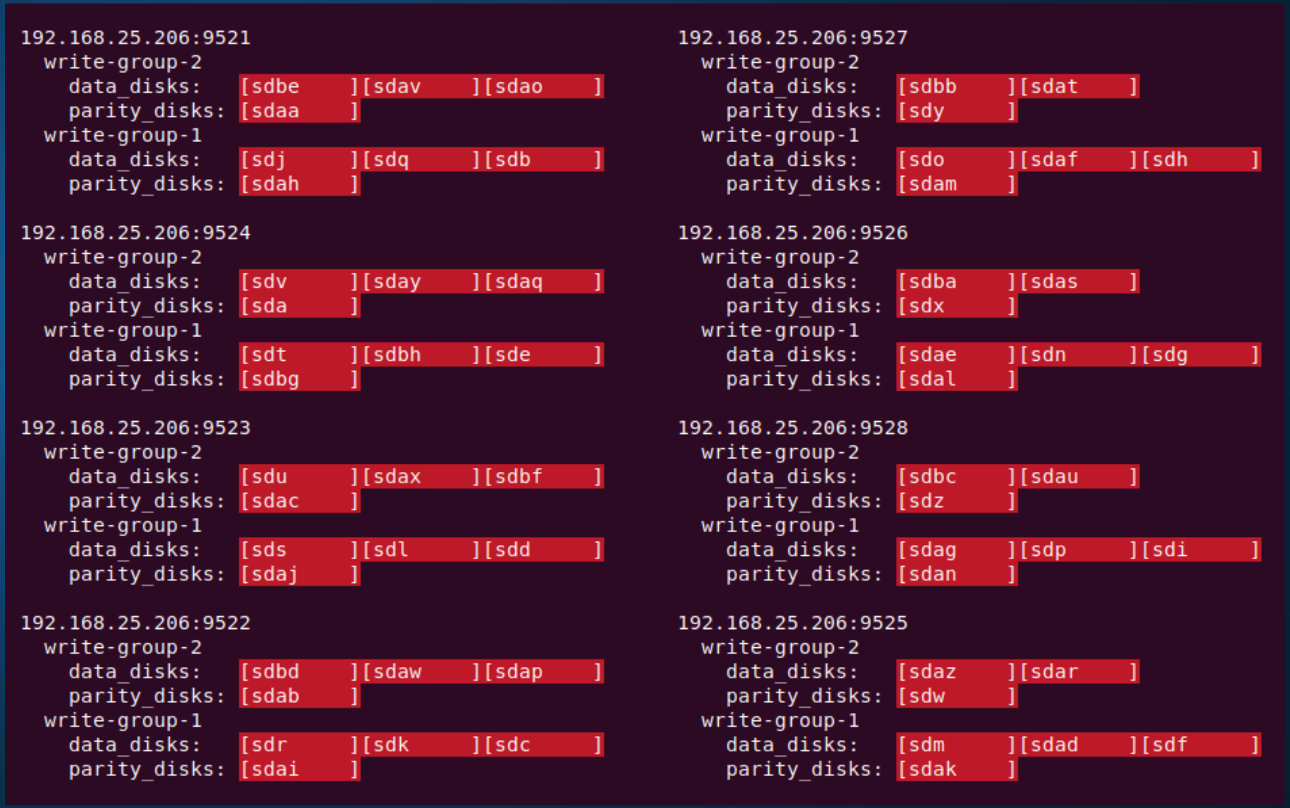
Testing Environment

Let's configure one physical server running multiple Chunkserver processes:

- 1 Western Digital Data 60 JBOD.
- 60 HM-SMR Drives (28 TB each).
- 8 Chunkserver processes for EC(6,2): 4 processes with 8 drives and 4 processes with 7 drives, totaling 60 drives.
- 2 Write groups for each Chunkserver process.

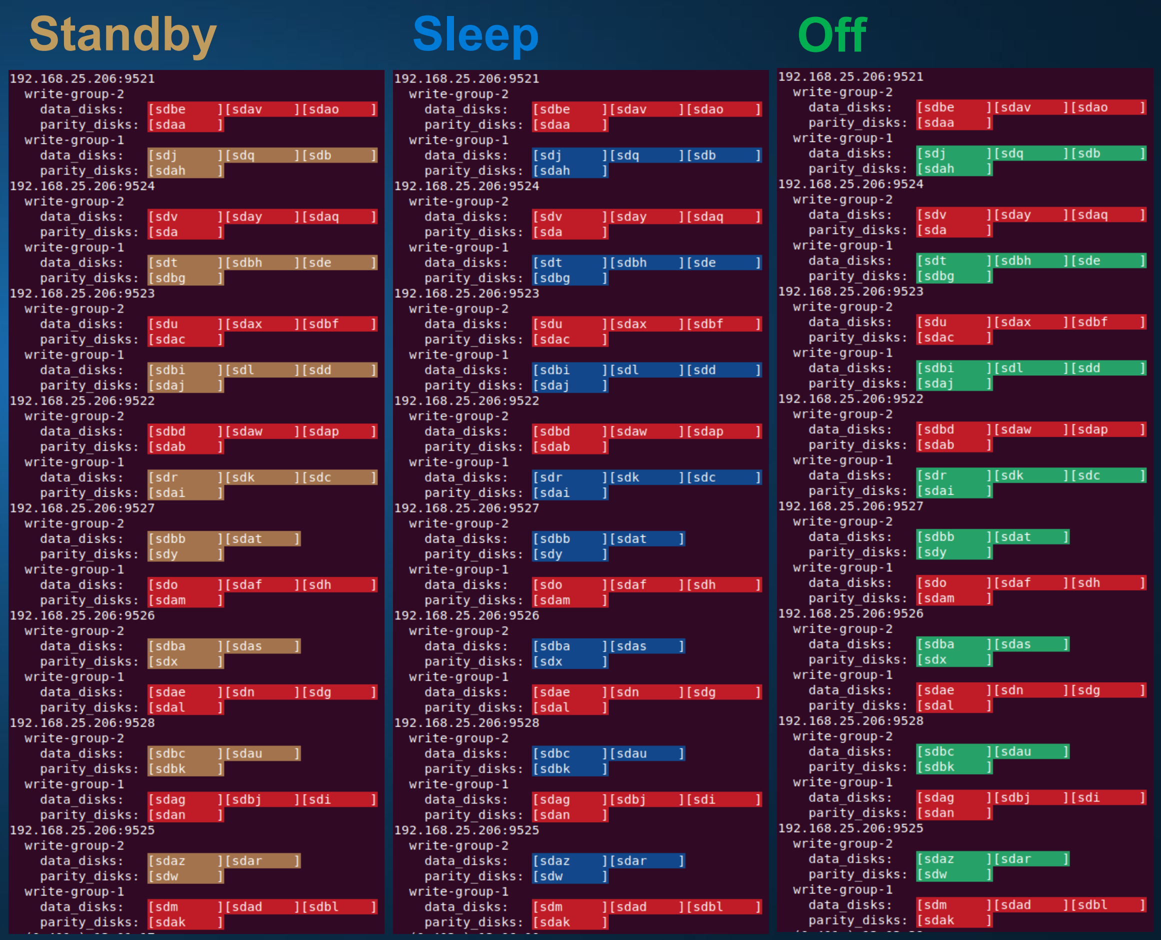
Drive Power States

- Without Write Grouping, all the drives remain in Active/Idle state (red) most of the time.
- Even if the timers to transition the disk between power states are well configured.

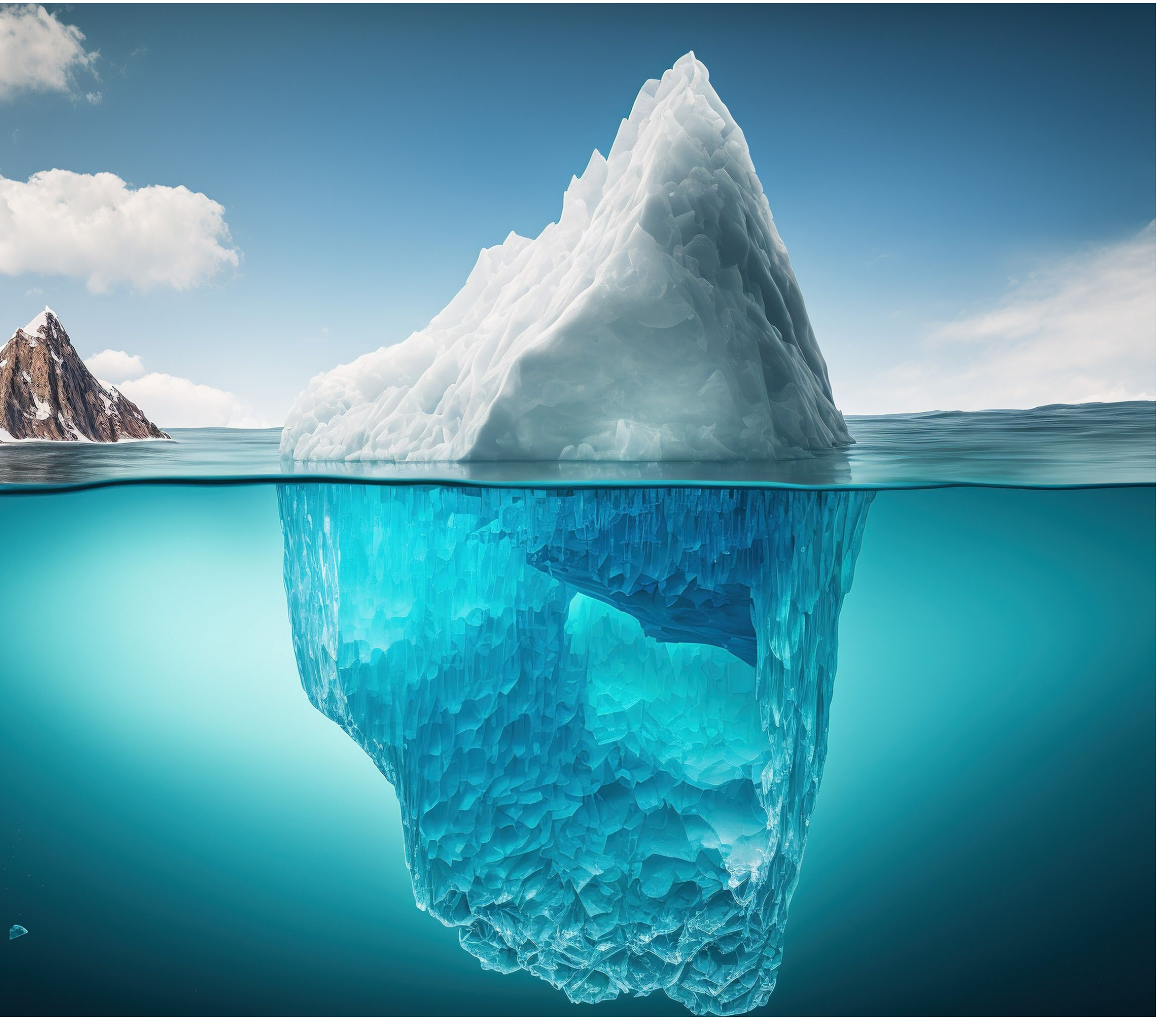


Drive Power States

- Only the active group remains in Active/Idle most of the time.
- With inactive groups protected from unwanted IO, the drives can transition to power friendly states.
- And stay in those states for longer periods.
- Cold drives can even be powered off.



Conclusions

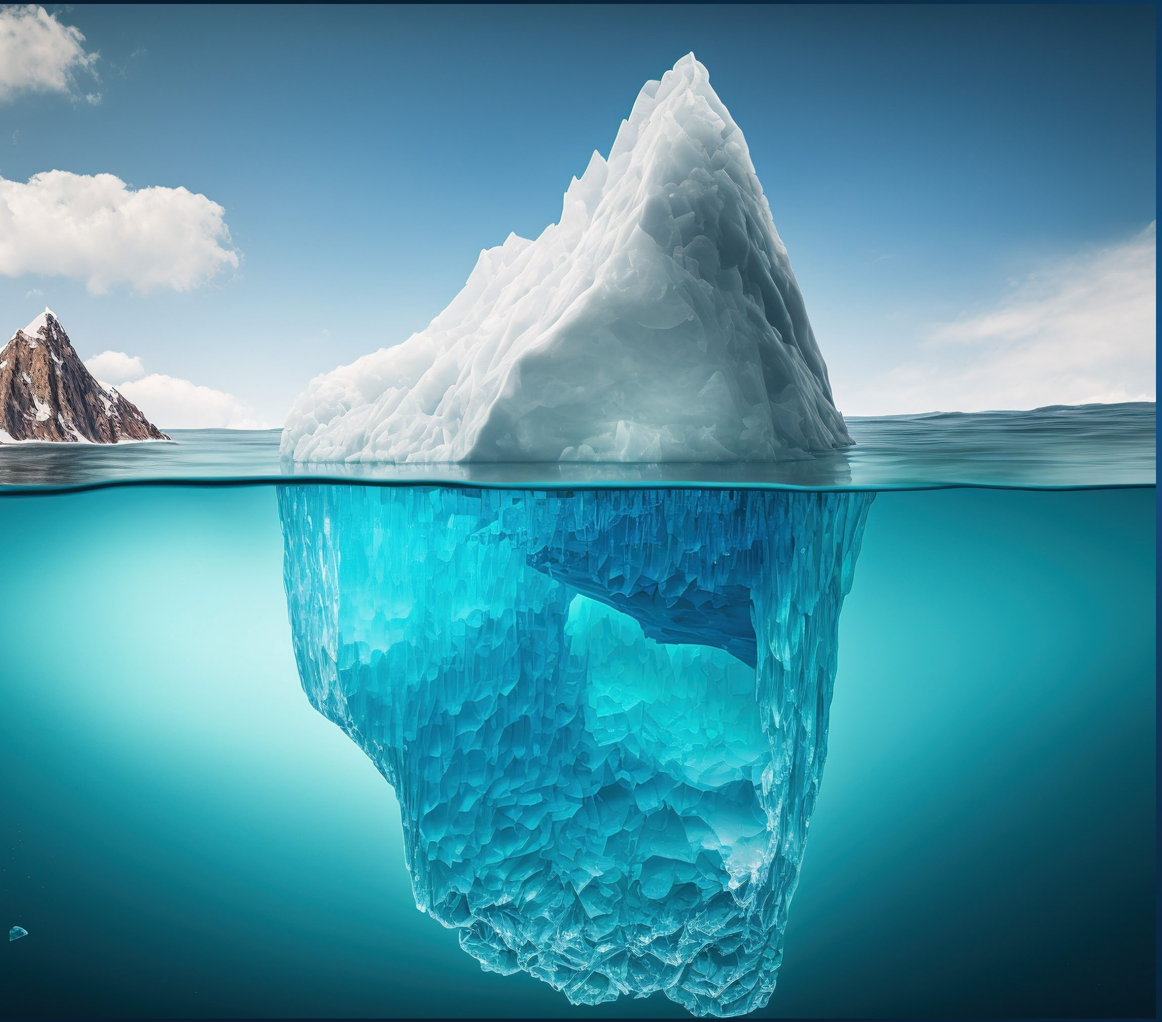


- We introduced the Infinite Cold Engine (ICE) to enhance energy efficiency in large-scale storage systems.
- The first phase is built on Write Grouping, which combines Workload Skew and Dynamic Power Management.
- The Diagonal Algorithm accelerates rebalancing during capacity expansion.
- These concepts have been successfully implemented in the SaunaFS distributed file system.

What’s Next?

Infinite Cold Engine (ICE)

HDD Power-off: Why Not MAID?



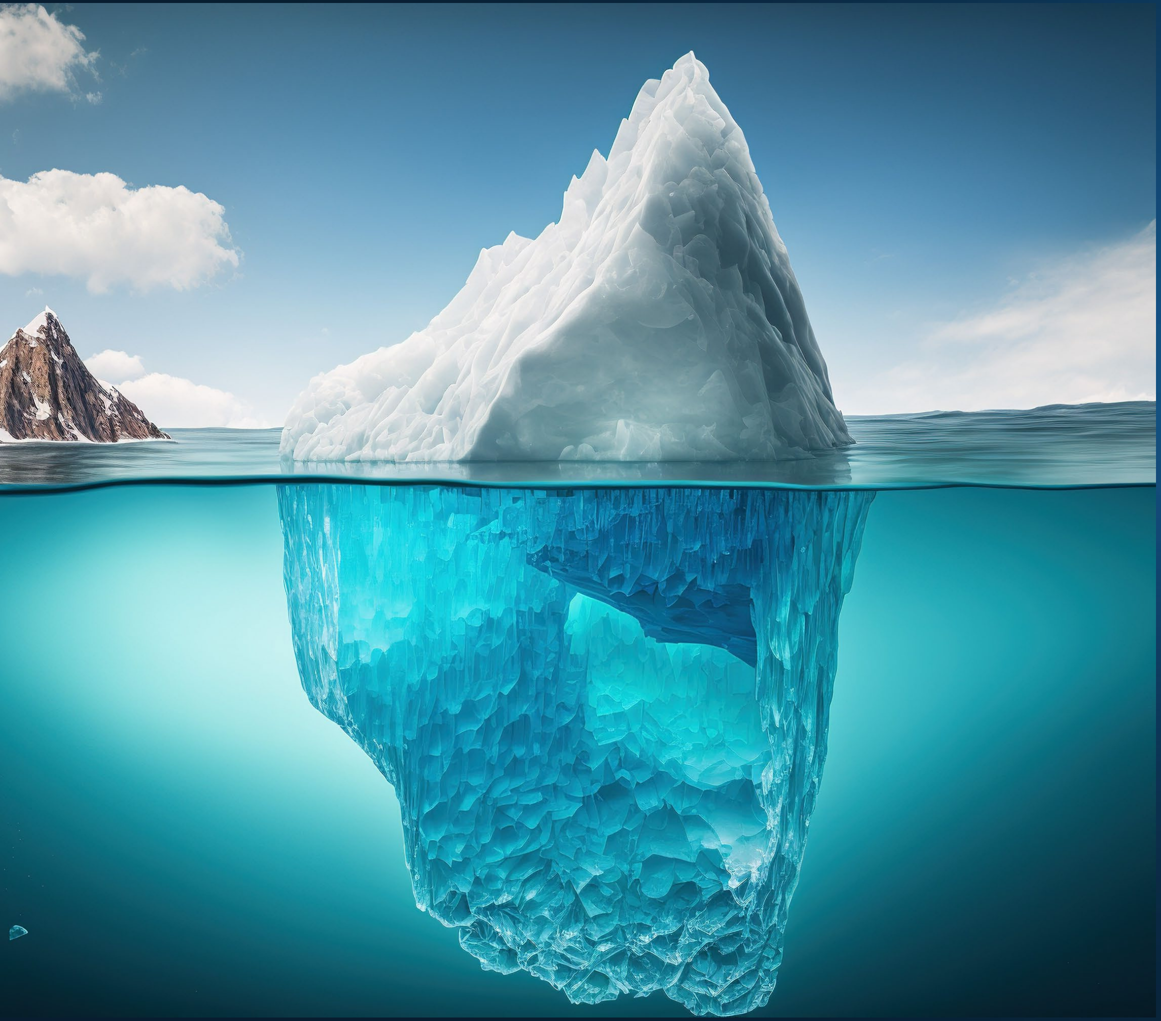
Data Distribution Challenge:

- Traditional storage solutions distribute data evenly, making it hard to power down drives

Why no broader adoption happened:

- No scale-out (scale-up)
- No software-defined storage (HW raid)
- No workload focus (trying to support any workload)

ICE as Our Take on MAID (DPM + WS)



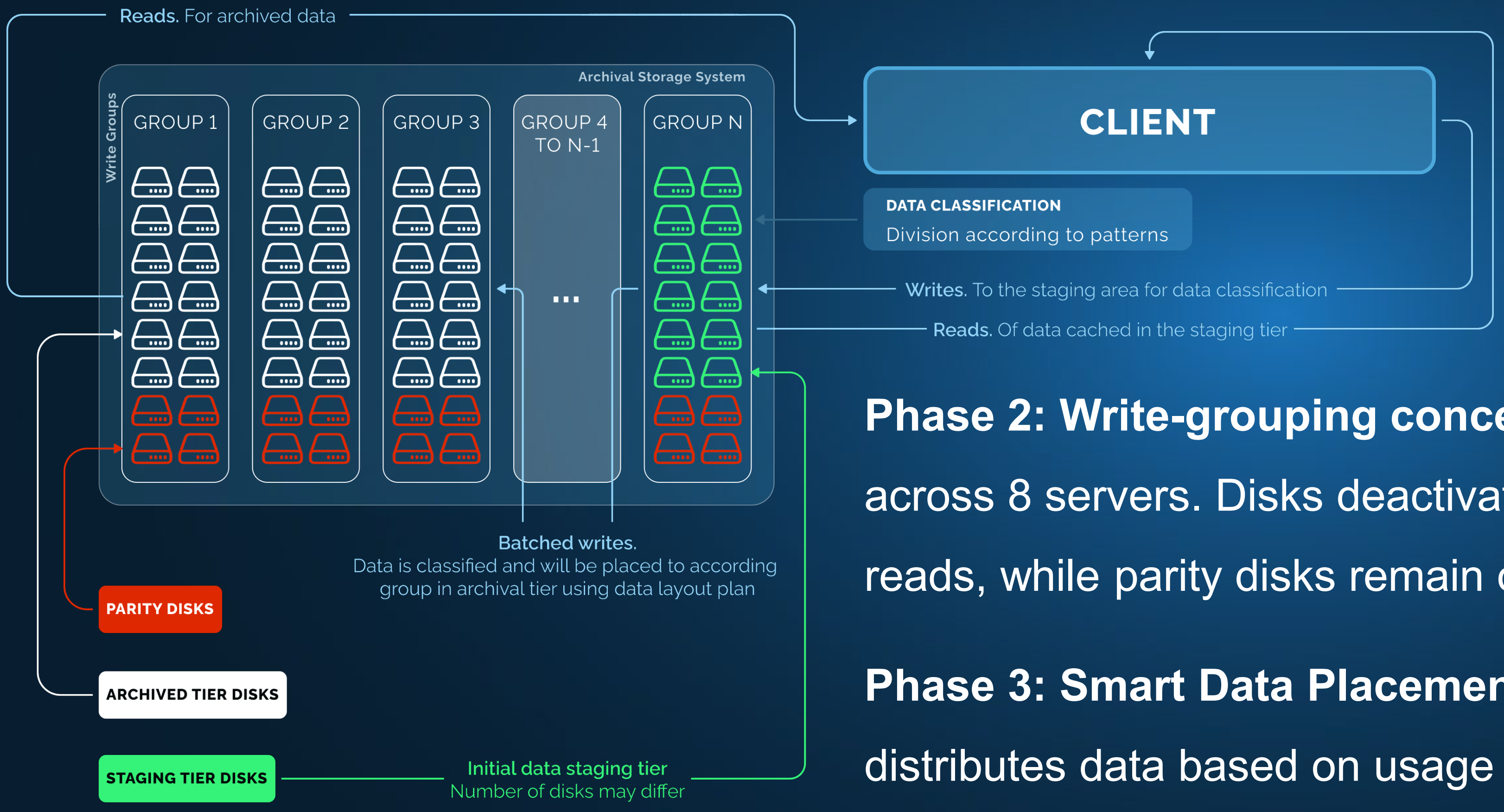
Infinite Cold Engine (ICE)

- Phase 1: HM-SMR support, **18% total energy savings**
- **Phase 2: Write-grouping conception Y2024 43%***
- Phase 3: Smart Data Placement for WG Y2025 **50%***
- Phase 4: AI-driven background service Y2026 **70%***

* Projected total energy savings

SDC 24

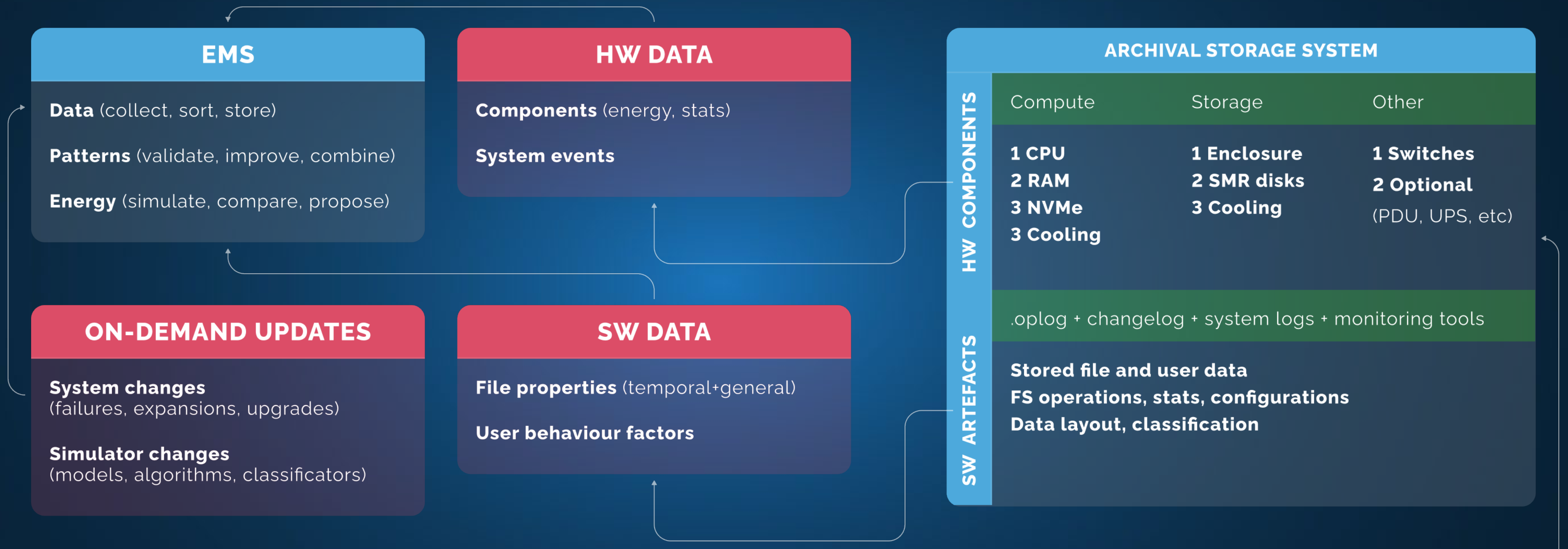
ICE: Combining Write-Grouping with SDP

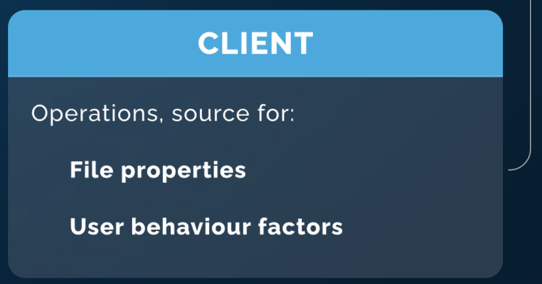


Phase 2: Write-grouping conception. SMR disks are grouped across 8 servers. Disks deactivate when full, reactivating only for reads, while parity disks remain off.

Phase 3: Smart Data Placement for Write-Grouping. SDP distributes data based on usage patterns, first storing it on staging disks for classification, then moving it to main storage in batches.

AI-Driven Background Service





- Gather data on power usage, user behavior, and system changes.
- Test classification parameters through simulations.
- Compare real energy use with simulations and provide recommendations.

Thank you!

Your feedback is important to us.

Piotr Modrzyk

Principal Architect at Leil Storage and SaunaFS

pm@leil.io

David Gerstein

Founder & CTO at Leil Storage and SaunaFS

david@leil.io

