croit

CEPH DAYS CHICAGO · NOVEMBER 16, 2026

Ceph on a Shoestring and a Song:

Extended Director's Cut

Older hardware as a hedge against rising component prices

François Baillargeon · Ceph & Proxmox Support, croit

ABOUT ME

François Baillargeon

Storage & Virtualization , Croit



Ceph, Proxmox, and 26 years of IT, helping teams make the best of their dollar.

Storage & Ceph

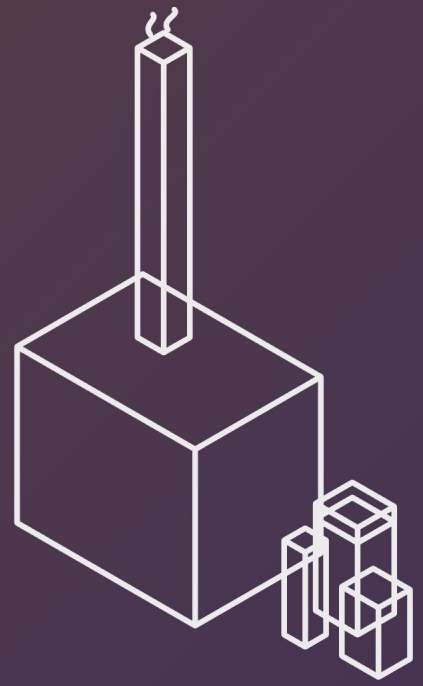
Proxmox & Virtualization

Automation & DevOps

Based in Quebec City, originally from Rouyn-Noranda, Québec, Canada.

WHY THIS MATTERS TO ME

A Reason Closer to Home



I'm from Rouyn-Noranda, Québec (Canada), home to the Horne Smelter, North America's largest metals recycler.

It recycles the metals out of old electronics, including servers, but decades of emissions have left the local population with elevated blood arsenic and lead levels.

Reusing hardware longer isn't just a budget hedge. It's fewer boxes that ever need recycling.

THE PROBLEM

The Price of New

DRAM is up over 250% since last October. Flash has more than doubled. AI-scale demand is eating the same supply chain.

Flash & SSD

More than double the price of a year ago.

Memory

DRAM is up over 250% since last October.

New Platforms

Current-gen CPUs, NICs, and chassis pass that cost to you.

Same acquisition budget. Fewer new nodes than it used to buy.

THE IDEA

Same Budget, More Nodes

Ceph doesn't care how old a node is. CRUSH only cares how many failure domains you have.

The lever Ceph hands you is **host count**, not clock speed.



fewer, newer → more, older: 5, 8, even 10 years old

Redundancy, rebalancing headroom, and EC math are all functions of host count.

There is such a thing as fast enough.

UNDER THE HOOD

How CRUSH Places Data

Object
Its name is hashed to pick a placement group.

→

Placement group
A pool is split into PGs. Each PG is placed and recovered as one unit.

→

CRUSH rule
Walks the map and picks OSDs, one per failure domain.

→

OSDs
Replica 3 lands on 3 hosts. EC 8+3 lands on 11.

Clients compute placement themselves from the CRUSH map. There is no lookup table and no central server in the data path.

CRUSH asks for distinct hosts, not for fast ones.

UNDER THE HOOD

Failure Domains Are the Real Currency

OSD
one drive

→

Host
one server

→

Rack
shared power and switch

→

Room
one room or site

The default rule: one per host

Every copy or chunk goes to a different host, so a dead server never takes two of them.

Small clusters feel it first

With 6 hosts and EC 4+2, every host holds a chunk of every PG. There is no spare host to rebuild onto.

You are buying failure domains. The silicon comes along.

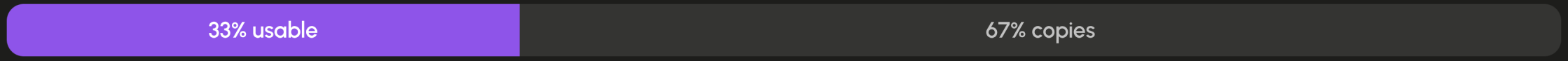
THE BASELINE

Where Replication 3 Tops Out

Replica 3 is the sensible default: simple and safe, even on a handful of hosts.

33%
usable of raw capacity

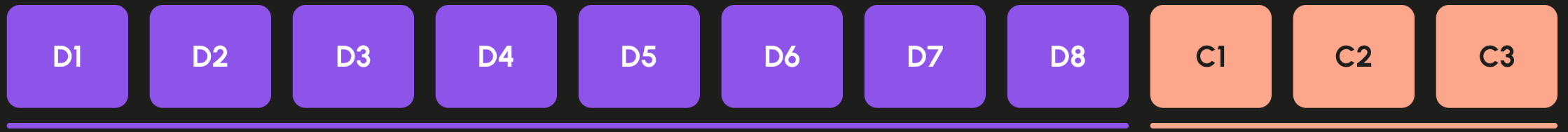
2
failures tolerated at once



That ratio never improves, no matter how many hosts you add.

UNDER THE HOOD

How Erasure Coding Works



usable = k / (k + m) = 8 / 11 = 73%

Any 8 of the 11 chunks are enough to rebuild the object.

Each chunk is stored on a different host.

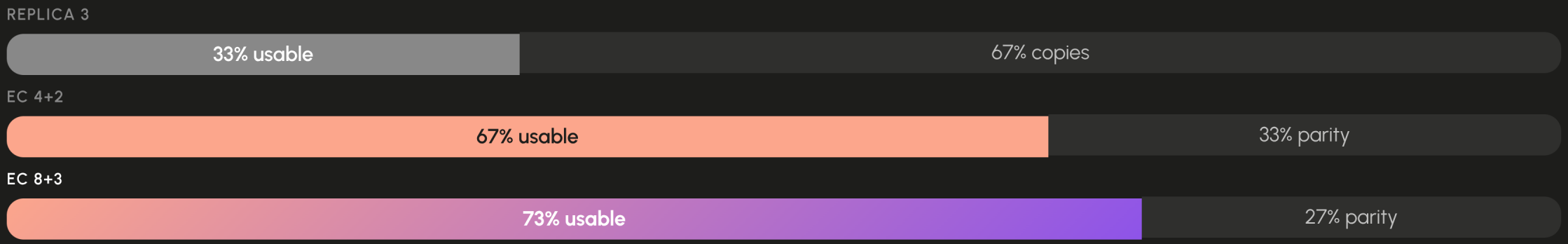
Raw cost is 1.375× the data, against 3× for replica 3.

THE UNLOCK

Node Count Unlocks Erasure Coding

Scheme	Efficiency	Failures	Hosts needed
Replica 3	33%	2 at once	3 (4+ practical)
EC 4+2	67%	2 at once	6 (8–10 practical)
EC 8+3	73%	3 at once	11 (16–22 practical)

USABLE CAPACITY, SIDE BY SIDE



UNDER THE HOOD

How Many Hosts Do You Actually Need?

k + m

The hard minimum

EC 8+3 needs 11 hosts just to place every chunk.

k + m + 1

The practical minimum

A spare host lets Ceph rebuild to full redundancy instead of staying degraded.

k + 1

Default min_size for EC

With 8+3, writes pause once fewer than 9 chunks are reachable.

For EC 8+3 that means 11 hosts at the floor and 16 to 22 in practice.

Count hosts first. Pick k and m second.

UNDER THE HOOD

Choosing k and m

Profile	Usable	Survives	Hosts (min / practical)
EC 4+2	67%	2 failures	6 / 8–10
EC 6+3	67%	3 failures	9 / 13–18
EC 8+2	80%	2 failures	10 / 15–20
EC 8+3	73%	3 failures	11 / 16–22
EC 8+4	67%	4 failures	12 / 18–24

Raise k

Better efficiency, but wider stripes, more rebuild reads and more hosts.

Raise m

More durability and more failures tolerated, at the cost of usable space.

UNDER THE HOOD

What a Rebuild Really Costs

REPLICA 3

1×

read per copy rebuilt

EC 8+3

8×

read per chunk rebuilt

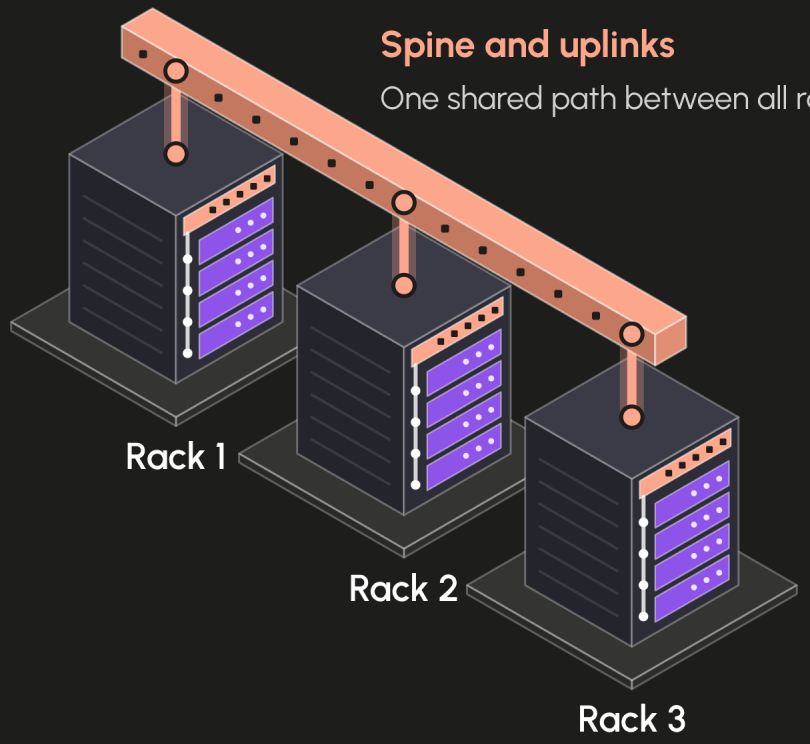
Plan the network for rebuilds: 10 or 25G, not 1G.

More hosts share the work, so each disk rebuilds less.

LRC and CLAY plugins trade a little space for cheaper repair.

UNDER THE HOOD

Fast Enough Has a Ceiling: The Uplink



A rack-level rule spreads chunks across racks, so each write crosses the uplinks.

A rebuild reads 8 chunks, and most of them travel between racks or rooms.

Hosts and disks can look idle while an oversubscribed uplink is saturated.

Check the oversubscription ratio, and keep the cluster network apart from the public one.

Consider multi-member LACP for increased speeds in complex CRUSH maps.

- Servers: plenty of capacity inside a rack
- Switches and uplinks: shared by all racks

Fast enough is set by the slowest link, not the oldest node.

UNDER THE HOOD

The Catch: Small Writes

Full stripes are cheap

Large objects and big sequential writes encode once and write all k + m chunks in one pass.

Good for object storage, backups and archives.

Small overwrites are not

Changing 4 KiB inside a stripe means reading the stripe, re-encoding parity and writing it back.

Painful for VM disks and databases.

RBD can sit on an EC data pool. Image metadata stays on a replicated pool.

Overwrites on EC pools need BlueStore, which every modern cluster runs.

The default stripe_unit is 4 KiB per chunk, so an 8+3 stripe is 32 KiB wide.

UNDER THE HOOD

Where EC Fits, Where Replica 3 Stays

Good fit for EC

Object storage (RGW)

Backups and archives

Large sequential files on CephFS

Cold and warm data

Keep on replica 3

VM disks and databases

Small random IO of any kind

Metadata pools: RBD, CephFS, RGW index

Anything where tail latency is the SLA

Most clusters end up running both.

Pools are chosen per workload, so mix them freely on the same hardware.

THE TRADE-OFFS

What Actually Costs You

CPU headroom

EC costs more CPU than replication, especially during recovery.

Network

Rebuilding a shard reads *k* others: plan real 10/25/100G.

Drive age

Used enterprise drives often have 80–95% life left: verify with SMART, don't assume.

Power & cooling

Older platforms draw more watts per usable terabyte.

Failure domains

CRUSH needs hosts on separate PDUs, switches, racks.

Firmware & support

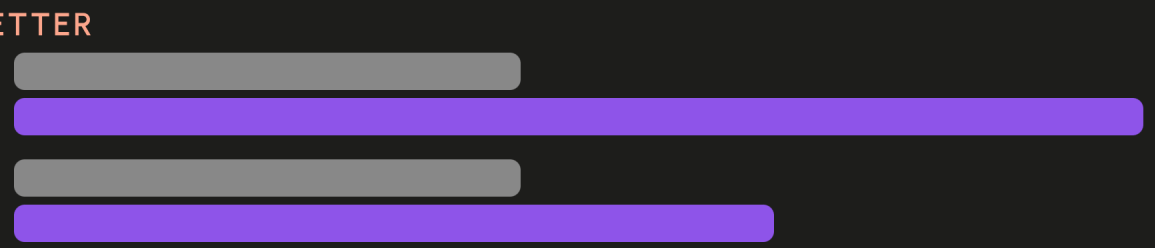
Older platforms may be past vendor security updates.

FLASHSTORE ON REPLICATED POOLS

Replicated Pools: Up to 2.2× the Throughput

Replica 3, 30 NVMe OSDs, 48 clients, against stock BlueStore on Ceph 21.2.0.

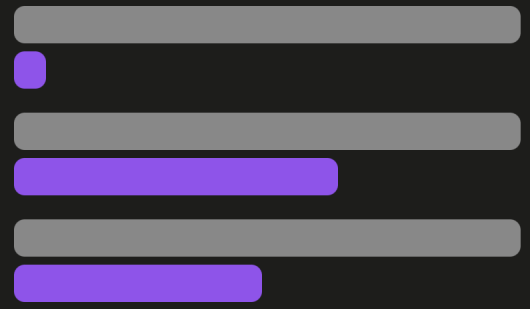
MORE DONE · LONGER BAR IS BETTER



2.2×
0.99M → 2.21M IOPS

+50%
4K random write: 2.1×

LESS SPENT · SHORTER BAR IS BETTER



16× lower
12.3 → 0.76 ms, avg -52%

-36%
best case -59%

-51%
5.3 → 2.6 GiB, with cache

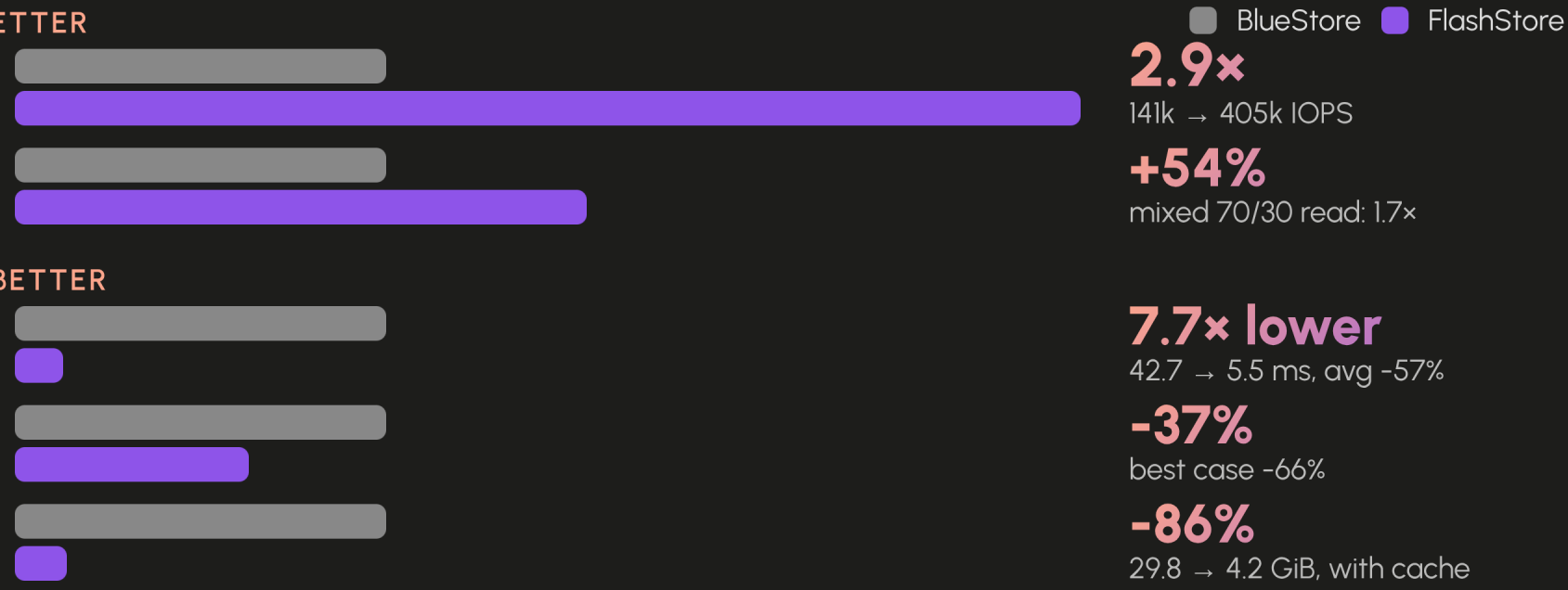
Same hardware, up to 2.2× the throughput, less CPU per op.

Why: one pinned thread per core, a COW B+tree instead of RocksDB, one copy per write.

Rows marked best case are single cells, the rest are averages. Single run per test. Preview, not yet recommended for production data.

Erasure Coded Pools: Up to 2.9× the Throughput

Erasure coded 2+2, 30 NVMe OSDs, 48 clients, against stock BlueStore on Ceph 21.2.0.



Small writes hurt EC most. Here they run up to 2.9× faster.

Why: BlueStore's RocksDB metadata cache balloons on small EC writes. FlashStore uses direct I/O.

Rows marked best case are single cells, the rest are averages. Single run per test. Preview, not yet recommended for production data.

FLASHSTORE

The Mad Wizards



Martin Verges



Md Mahamudur Rahaman Sajib



Igor Fedotov

IN PRACTICE · ILLUSTRATIVE

A Worked Example

ROUTE A — FEWER, NEWER

6 new hosts



All new, current-gen. Fine for replica 3, not enough hosts for EC 8+3.

33% usable

ROUTE B — MORE, OLDER

18 off-lease hosts



Same budget, spent on off-lease previous-gen gear, plenty of headroom for EC 8+3.

73% usable

Same budget. ~2.2× the usable capacity, one more failure tolerated.

Illustrative: your ratios will vary by chassis and disk choice

RUNNING IT

Headroom: Plan for the Host You'll Lose

max fill ≈ 85% × (N – 1) / N

85% is Ceph's default nearfull warning. Assumes equal hosts that share a rebuild evenly.

ROUTE A · 6 HOSTS

71%

safe ceiling if one host can fail

ROUTE B · 18 HOSTS

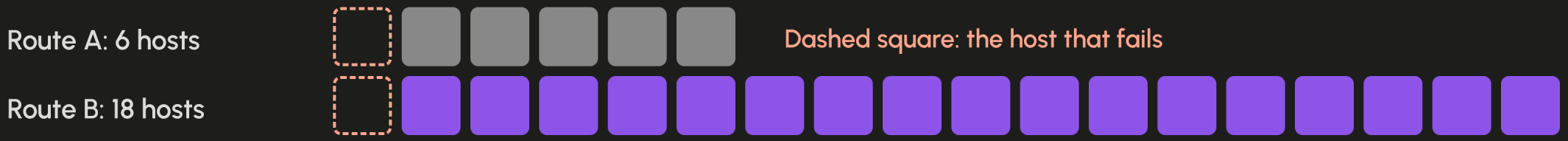
80%

safe ceiling if one host can fail

More hosts, more of the disks you paid for stay in use.

RUNNING IT

Smaller Failure, Faster Recovery



1 OF 6 HOSTS FAILS

16.7%

of the data to rebuild, about 3.3% on each of 5 survivors

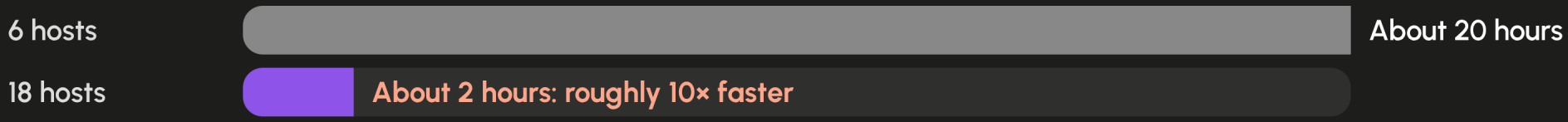
1 OF 18 HOSTS FAILS

5.6%

of the data to rebuild, about 0.33% on each of 17 survivors

TIME TO FULL REDUNDANCY AFTER ONE HOST FAILS

Illustrative: same total data, same recovery rate per host



Same budget, many small failures instead of a few big ones.

RUNNING IT

Mixed Generations in One CRUSH Map

Weights follow capacity

CRUSH weight tracks drive size, so a 4 TB drive gets half the data of an 8 TB one.

Device classes

Tag OSDs as hdd, ssd or nvme and write rules per class, so media never mix in one pool.

Primary affinity

Keep slow or aging OSDs from acting as primary for reads on replicated pools.

Stagger generations across racks so one bad batch never owns a whole failure domain.

RUNNING IT

Running Old Hardware Safely

Watch SMART closely

Alert on reallocated and pending sectors before a drive fails outright.

Throttle recovery

Use the mClock profile, or cap backfills, so rebuilds never starve client IO.

Scrub on a schedule

Deep scrubs catch silent bit rot. Old drives need them more, so run them off-peak.

Balance with upmap

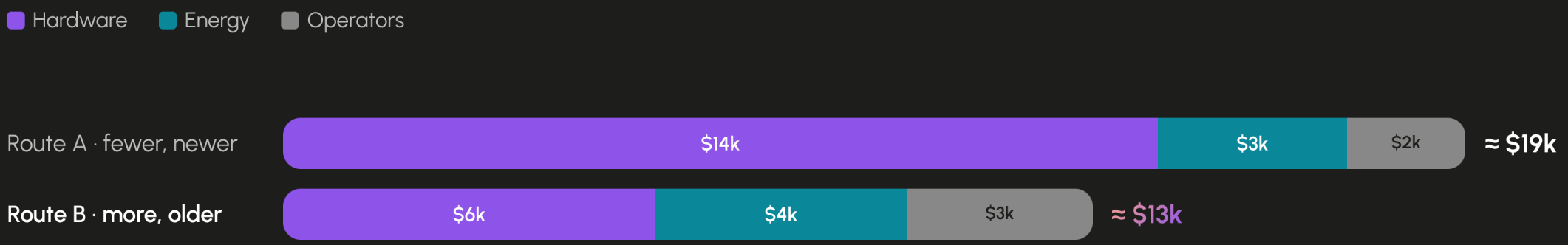
The upmap balancer evens out PG counts so old nodes are not quietly overloaded.

Old hardware is fine. Unwatched hardware is not.

THE REAL METRIC

Total Cost Per Petabyte-Year

Same illustrative fleet as before: hardware, energy, and operators, priced per usable petabyte per year.



Route B spends more on power and people, but it still wins on total, thanks to efficiency.

Spend on the people, not the things.

Illustrative: plug in your own $/kWh and labor rates

THE PLAYBOOK

Spend on Drives, Save on Chassis

Chassis secondhand, drives verified

Off-lease compute, retrofit with PCIe NVMe adapters.

Stagger generations

Mix hardware ages so failures and end-of-life don't cluster.

Size EC with headroom

Aim for total hosts around 1.5–2× (k+m).

Watch older nodes closer

SMART, temperature, and power draw become metrics.

TAKEAWAYS

Buy More Nodes, Not Just Bigger Ones

- Higher prices, same trick: Ceph only asks for failure domains, not new silicon.
- More hosts, even older ones, turns replica 3 into EC 8+3, better efficiency *and* durability.
- Spend the savings on drives, network, and monitoring, where reused hardware needs the help.

Thank you. Further reading: croit.io/blog · Building croit Ceph on Used Hardware