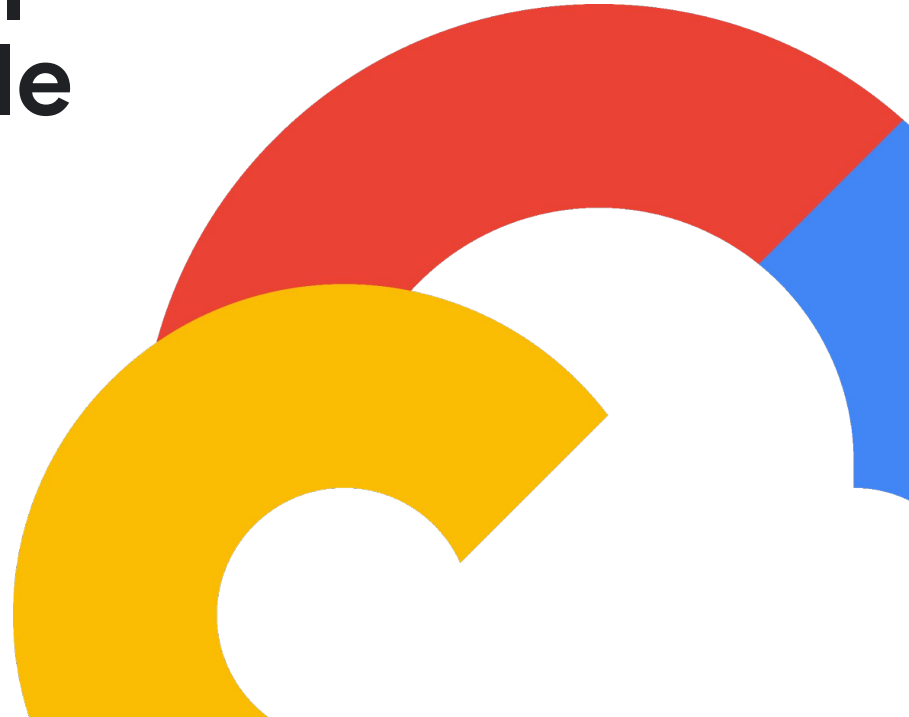


Google Cloud

Empowering capital
markets with Google
Cloud

Proprietary & Confidential

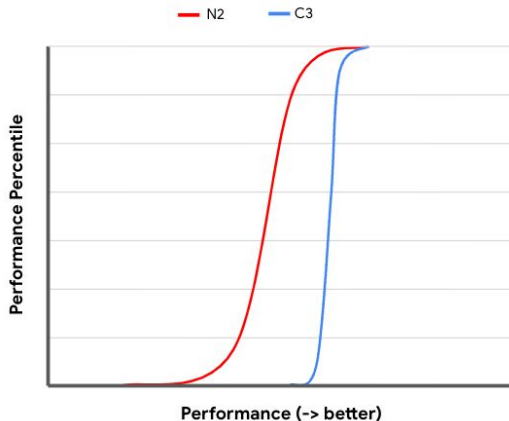
Google Cloud



Uncompromising performance consistency aligned to capital market needs

What differentiates C3 for capital markets?

- VMs strictly aligned to underlying NUMA
- Predefined VM shapes only (no custom machine types)
 - 4, 8, 22, **44, 88, 176 vCPUs** for optimal isolation
- Less frequent and disruptive maintenance events
- Adv. maintenance (28d frequency, 7d notification, controls)



Predictability

- Ensure applications and services run smoothly and meet their service level agreements.

Reliability

- Minimize downtime and unexpected failures due to maintenance events and noisy neighbors.

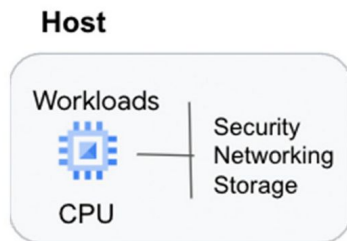
User Experience

- VM performance directly affects the end-user experience for match perf consistency

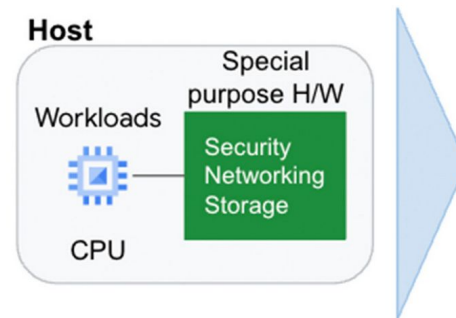
Resource Mgmt.

- Easier to manage and plan capacity and budgets for day-to-day needs and scaling events (e.g., holidays, weekends).

Titanium



Without offloads



Offloads on host only

- [Next](#) blog from 2023
 - [Titanium underpins Google's workload-optimized infrastructure | Google Cloud blog](#)
- DPDK extensions - ingress & egress benefits
- Checksum offloads
- Platform benefits (Google overhead offload, storage benefit, etc.)



<https://goo.gl/stac-titanium>

Aeron test environment

Infrastructure technology enablers

C3 machine optimizations

- NUMA architecture to alignment
- Simultaneous multithreading configuration SMT (eg. hyperthreading off)
- Placement policies (spread v. compact)
- 200 Gbps advanced networking

Network optimizations

- TCP/IP gVNIC offload functions
- DPDK for kernel bypass function
- Jumbo frame support

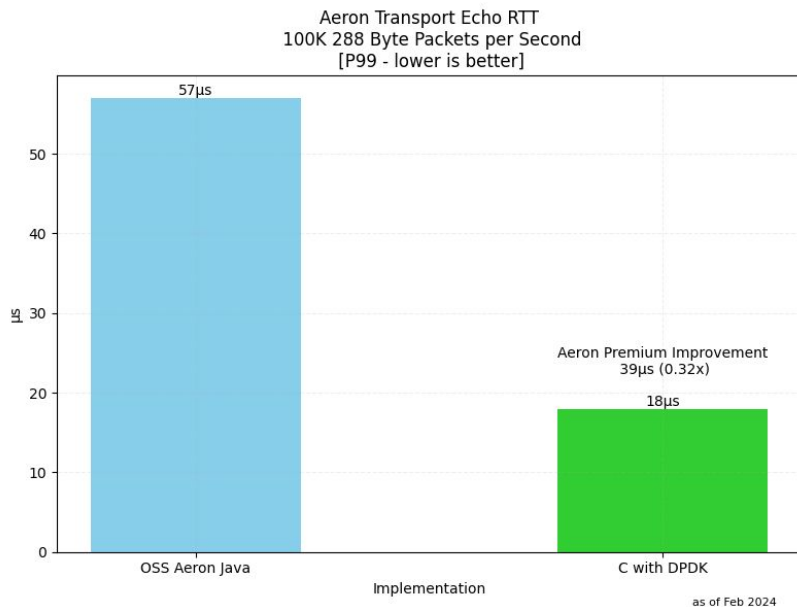
Google Aeron blog:
[Google Cloud optimized infrastructure for digital exchanges](https://goo.gle/aeron-stac)



<https://goo.gle/aeron-stac>

Open-source high performance messaging on Google Cloud*

Aeron Transport performance insights via Adaptive on Google Cloud's 3rd generation VM's



C3 machine type (C3-highmem-88)

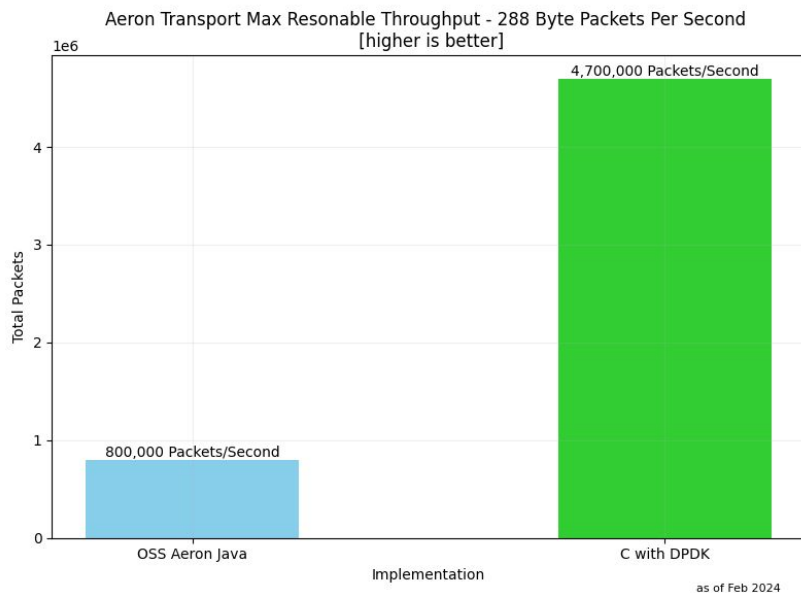
100K messages per second, 288 byte packet size
C with DPDK implementation

- P50: 15us (-50% v C2)
- P99: 18us (-60% v. C2)
- P999: 31us (-60% v. C2)



Open-source high performance messaging on Google Cloud*

Aeron Transport performance insights via Adaptive on Google Cloud's 3rd generation VM's

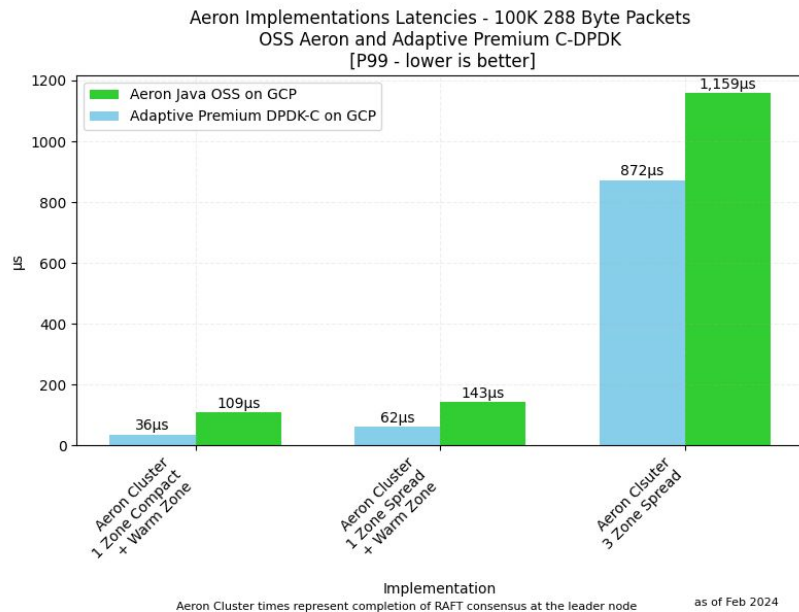


C3 machine type (C3-highmem-88) + DPDK
4.7M+ packets per second, 288 byte packet size –
10Gbps on a single thread
Current testing is indicating >9M packets per second



Open-source high performance messaging on Google Cloud*

Aeron Cluster performance insights via Adaptive on Google Cloud's 3rd generation VM's



C3 Machine Type (C3-highmem-88)

100K messages per second, 288 byte packet size
C with DPDK implementation

1 Zone Compact to 1 Zone Spread: <30µs latency
cost.

1 Zone HA to 3 Zone Spread: <1ms latency costs



Direct Peering

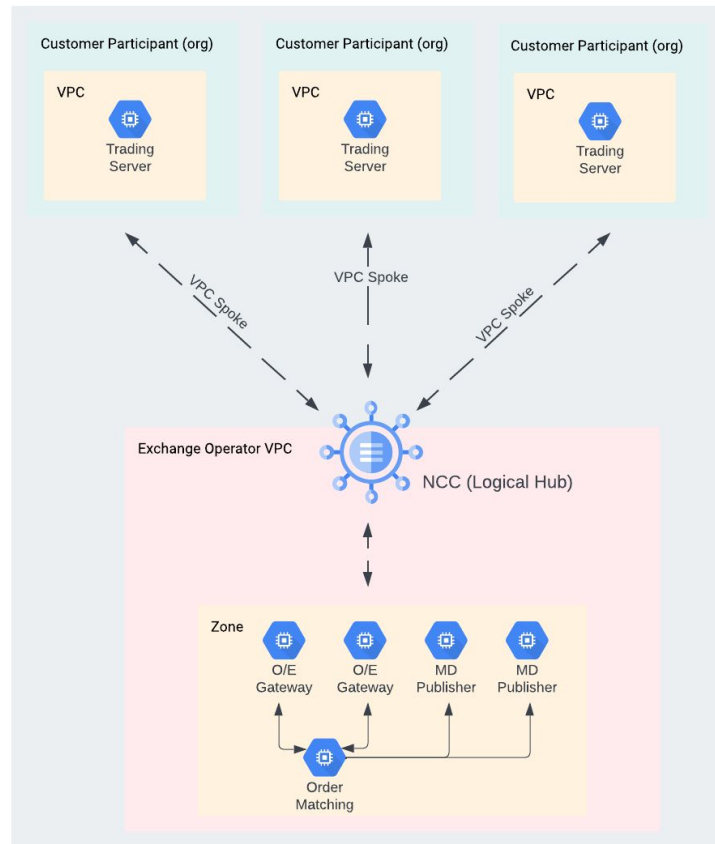
Enabling full participation and access to OMS/EMS/matching engines

Delivery patterns

- Full-mesh vs prefix filtering
- Interconnect support

Scale objectives

- 1000 organization target
- Direct Connect / Cloud Colocation pattern



Dynamic Workload Scheduler (DWS)

New obtainability capabilities for accelerators

Works across GCP

Managed instance
groups on GCE

Batch on
GCE

GKE

Vertex AI

Calendar Mode:

Job start times assurance
with future reservations

Use cases:
(re)training, recurring
fine-tuning

GPUs

Flex Start Mode:

Optimized economics and
higher obtainability for
on-demand resources

Use cases:
time flexible experiments,
fine tuning, batch inference

GPUs & TPUs

“

“The new DWS scheduling capabilities have been a game-changer in procuring sufficient GPU capacity for our training runs. We didn’t have to worry about wasting money on idle GPUs while refreshing the page hoping for sufficient compute resources to become available.”

**- Sahil Chopra, Co-Founder & CEO,
Linum AI**

Flex Start mode: AI/ML workloads get served in order of arrival

User scenario:

"I want to run my multi-node training job using 2 A2 VMs with 8 GPUs for 15h in us-central1"

Job parameters:

- Resource quantity (VM count)
- Location (Region or Zone)
- Run duration (default max 7 days)

Job queue - requests by arrival time



Capacity usage

