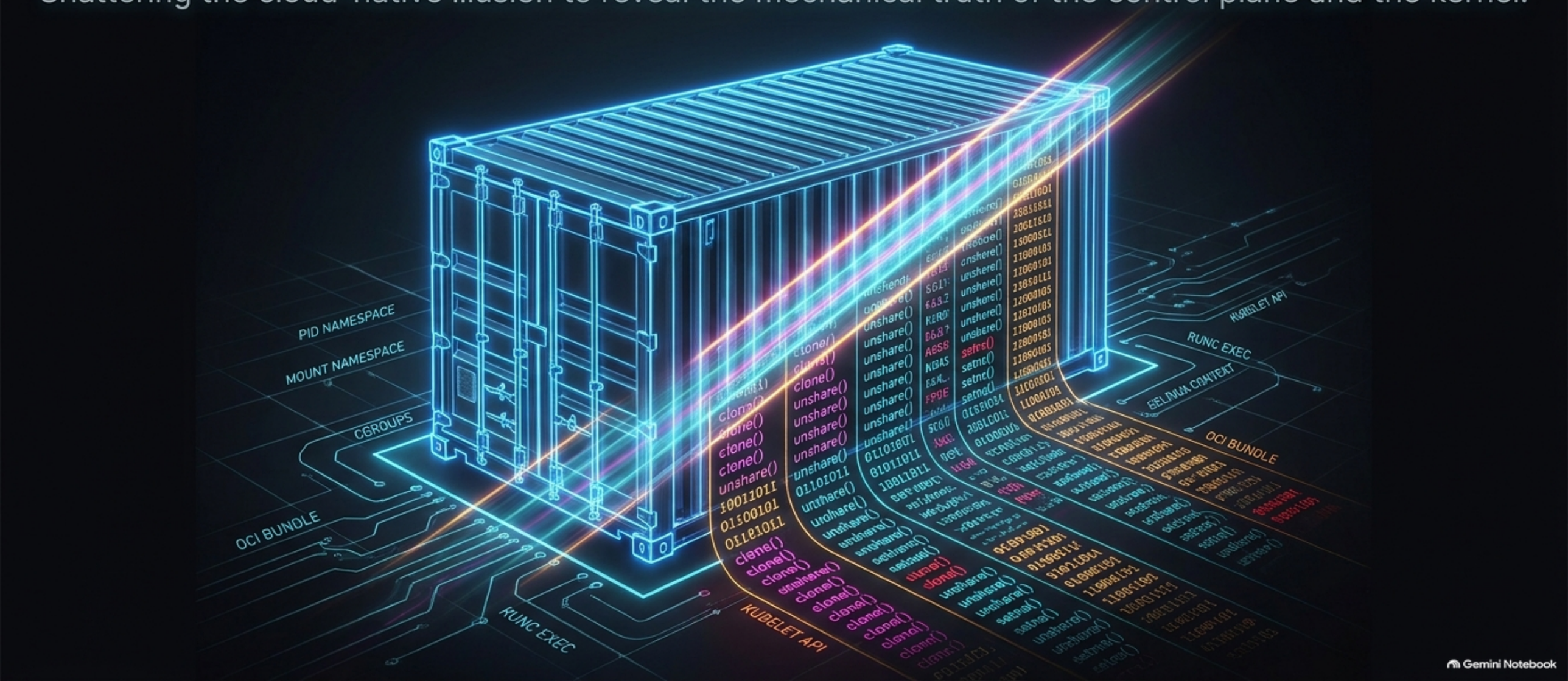
WEB ARCHITECTURE — S02E05: LINUX KERNEL CONTAINER PRIMITIVES, OCI RUNTIME SPECS & KUBERNETES ARCHITECTURE

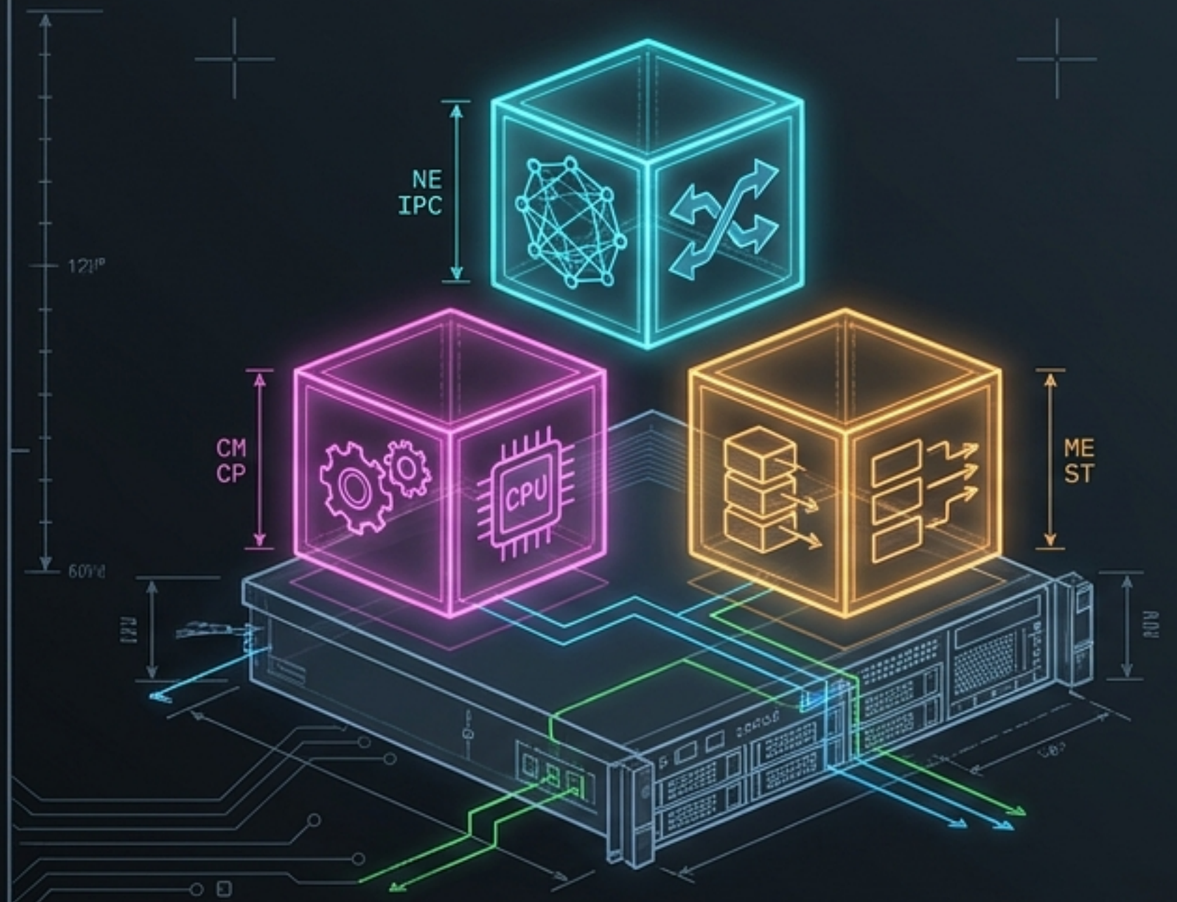
Shattering the cloud-native illusion to reveal the mechanical truth of the control plane and the kernel.



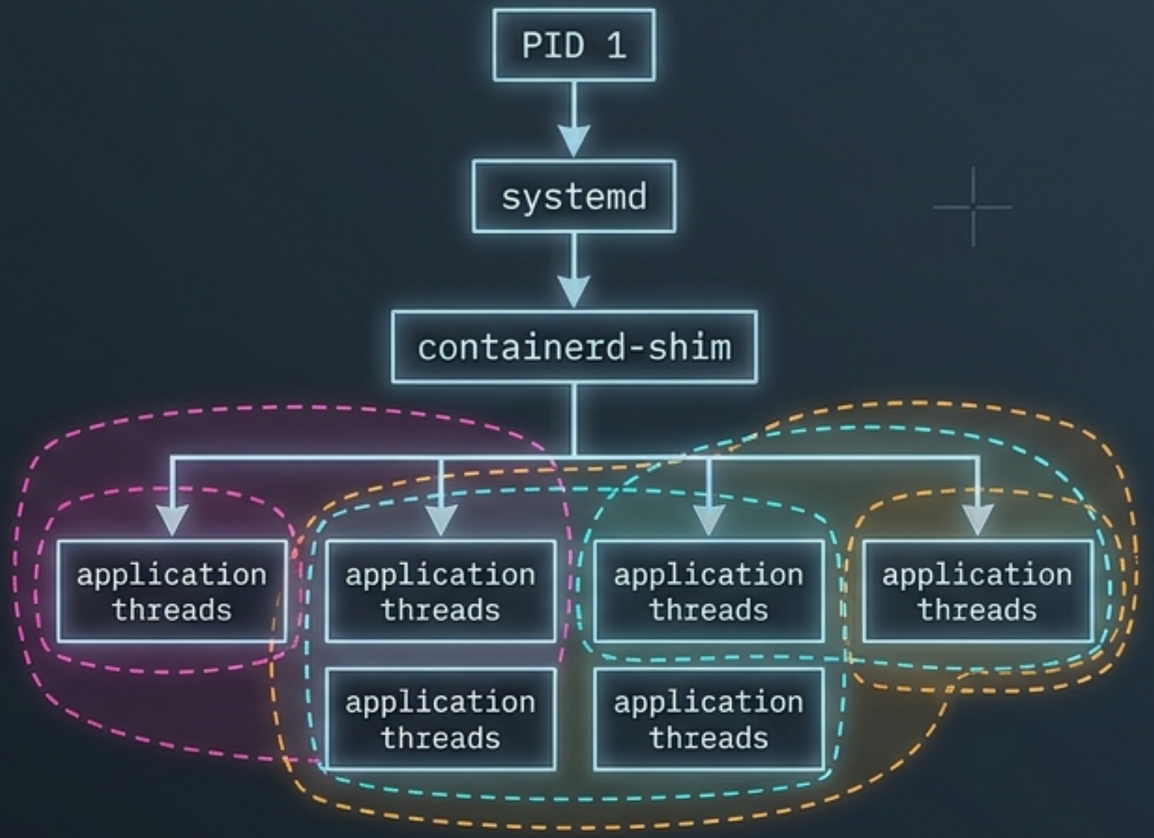
Gemini Notebook

The Grand Illusion: Containers Do Not Exist

The Illusion



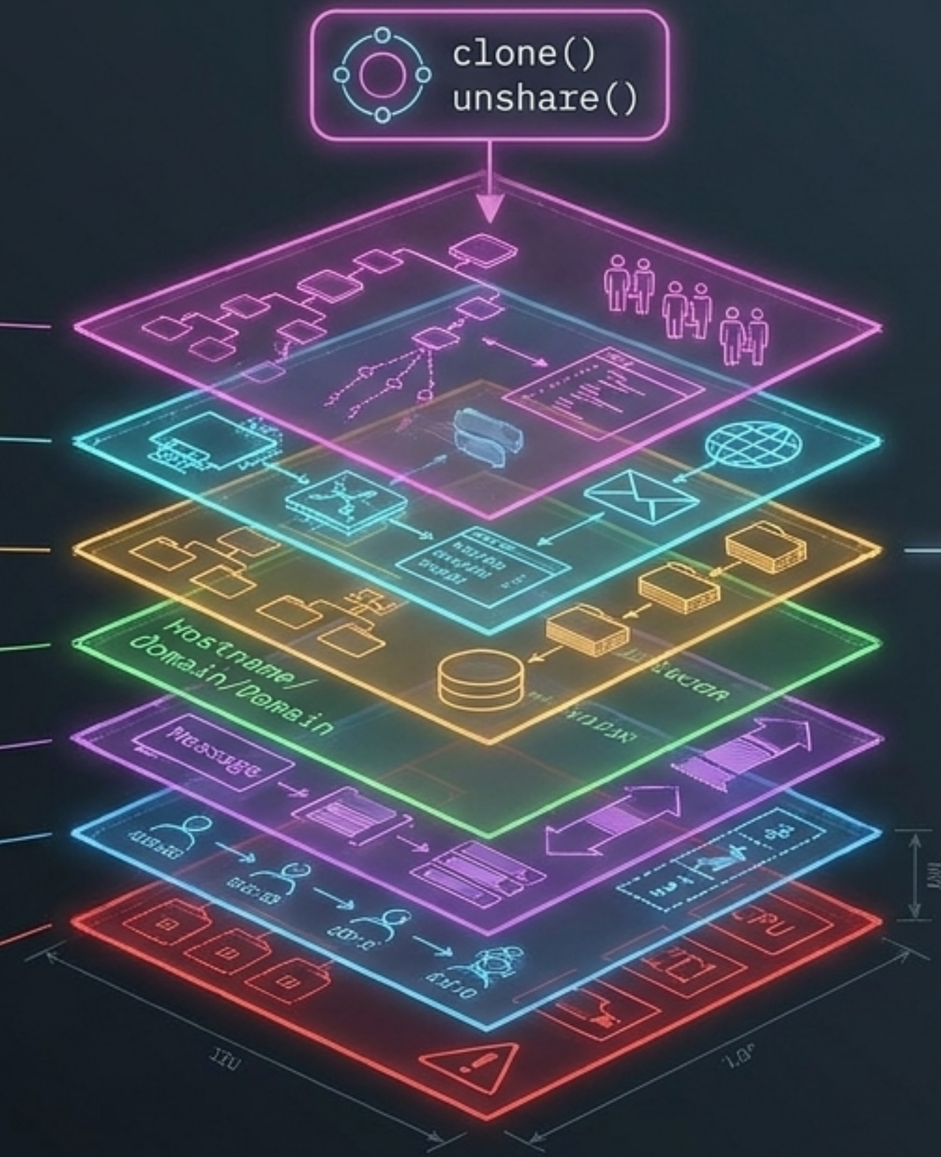
The Kernel Reality



"Containers" are not a first-class Linux kernel object. They are a user-space illusion constructed entirely from standard Linux processes subjected to strict view isolation and resource accounting.

Gemini Notebook

View Isolation: The 7 Linux Namespaces

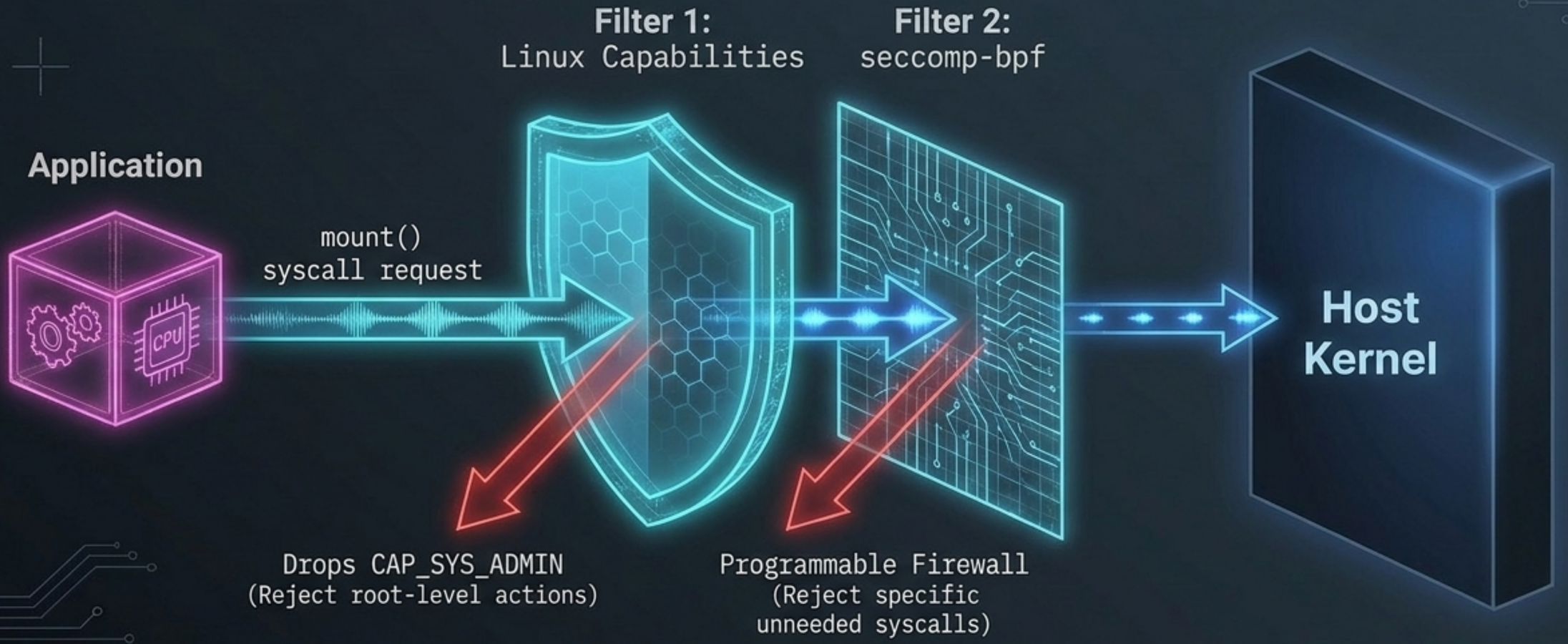


Namespaces wrap global system resources in an abstraction.

Monitored via /proc/pid/ns/ symbolic links, they ensure processes only see their own isolated instance of global resources.

Gemini Notebook

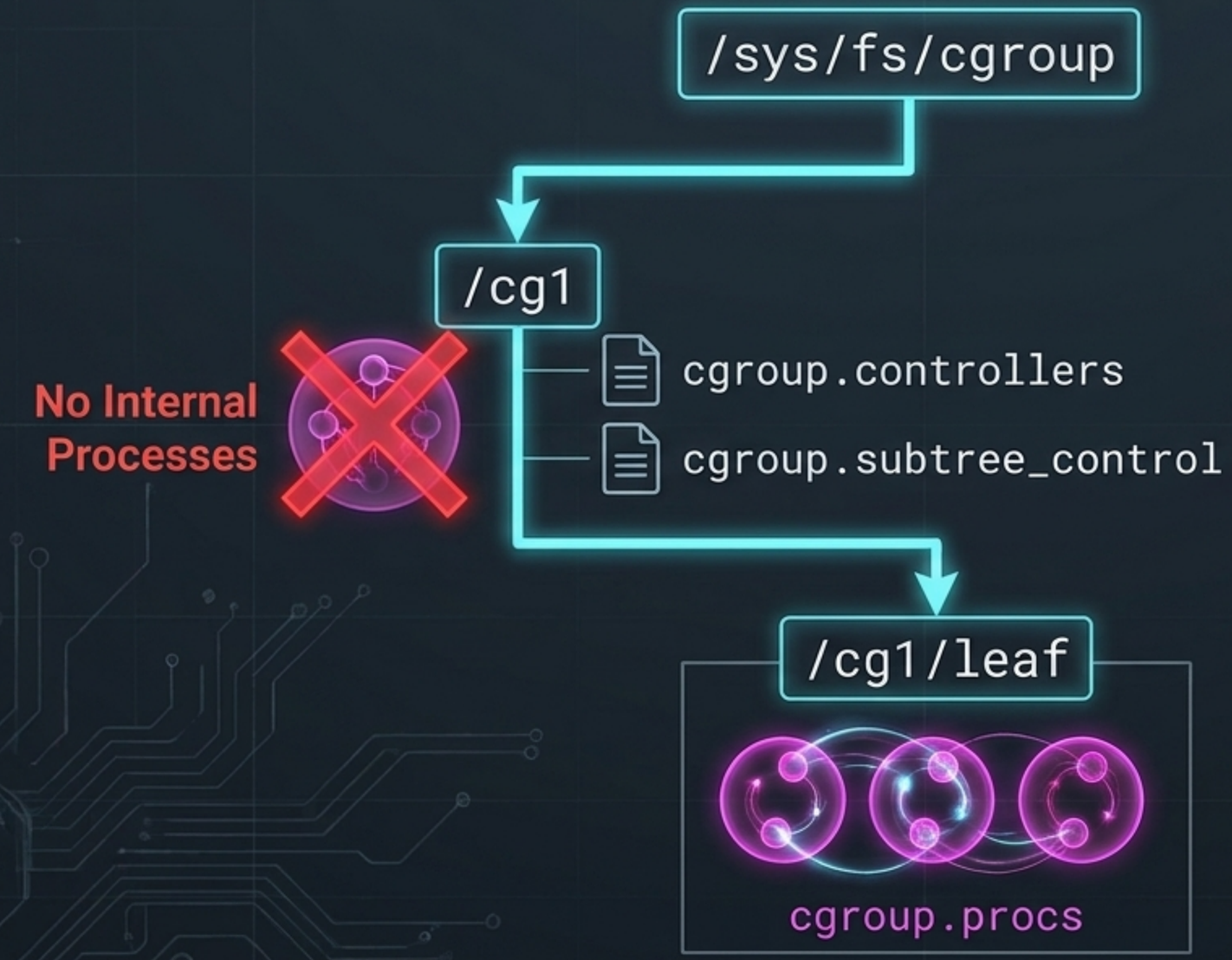
Syscall Lockdown: seccomp-bpf & Capabilities



Namespaces hide resources, but do not block system-level instructions. Secure isolation requires explicitly dropping Linux Capabilities and attaching eBPF filters (seccomp-bpf) to restrict the raw syscall surface area.

Gemini Notebook

Resource Governance: The cgroups v2 Unified Hierarchy

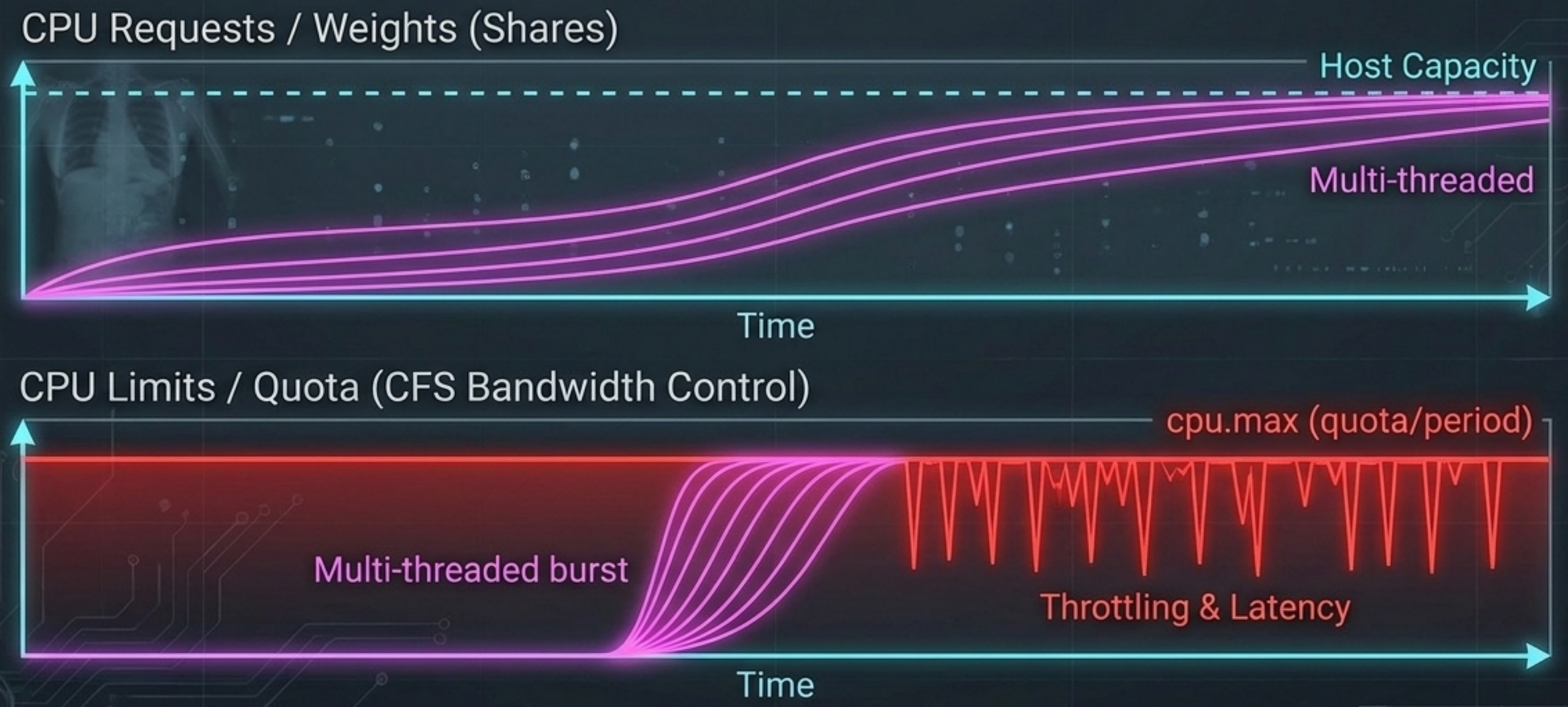


cgroups v2 resolves v1's inconsistencies by enforcing a **single unified hierarchy** and a strict "no internal processes" rule.

Resources are passed down through `cgroup.subtree_control`, ensuring predictable, unambiguous limits.

Gemini Notebook

CPU Requests vs. Limits: The CFS Throttling Trap



CPU Limits rely on CFS bandwidth control (`cpu.max`). Multi-threaded runtimes can consume a `100ms quota` in 10ms by using 10 threads, causing artificial throttling for the remaining 90ms.

Gemini Notebook

Memory Governance & The Kernel OOM Killer

Memory is an incompressible resource.

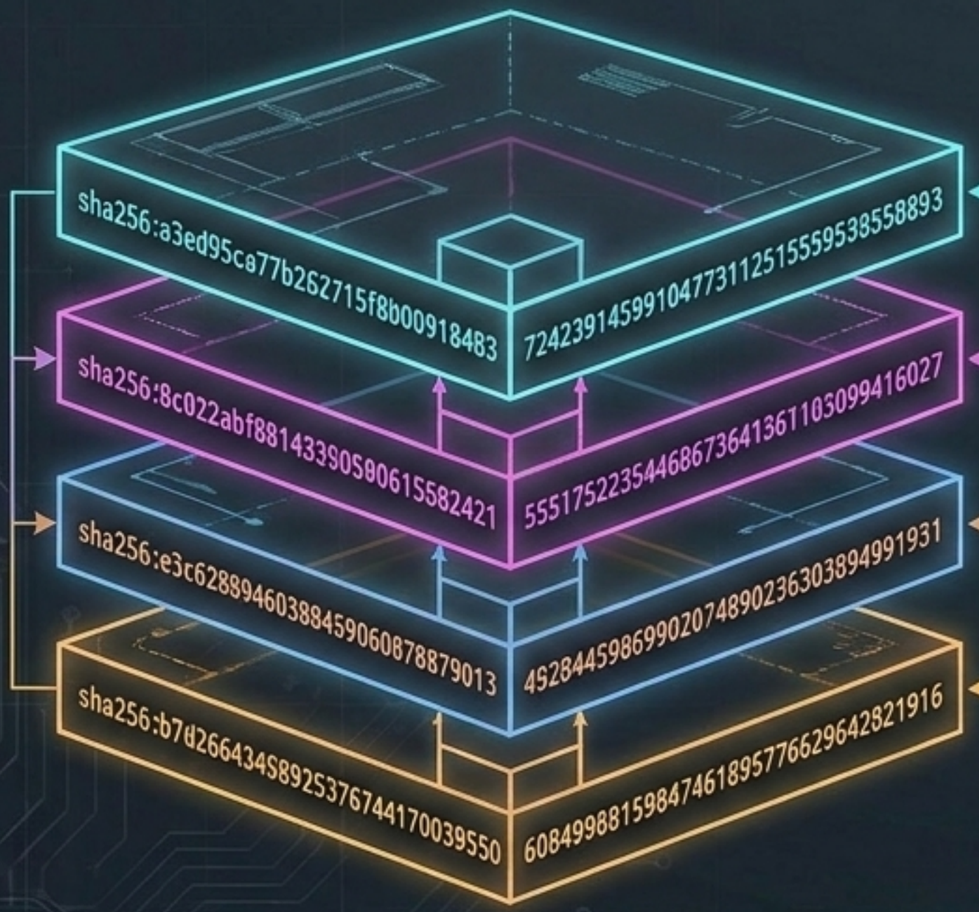
Exceeding the `memory.max` boundary triggers the kernel's **Out-Of-Memory (OOM) Killer**.

The kernel intervenes without warning, instantly terminating the process to protect host stability.



Gemini Notebook

The OCI Image Specification: Not Magic, Just Tarballs

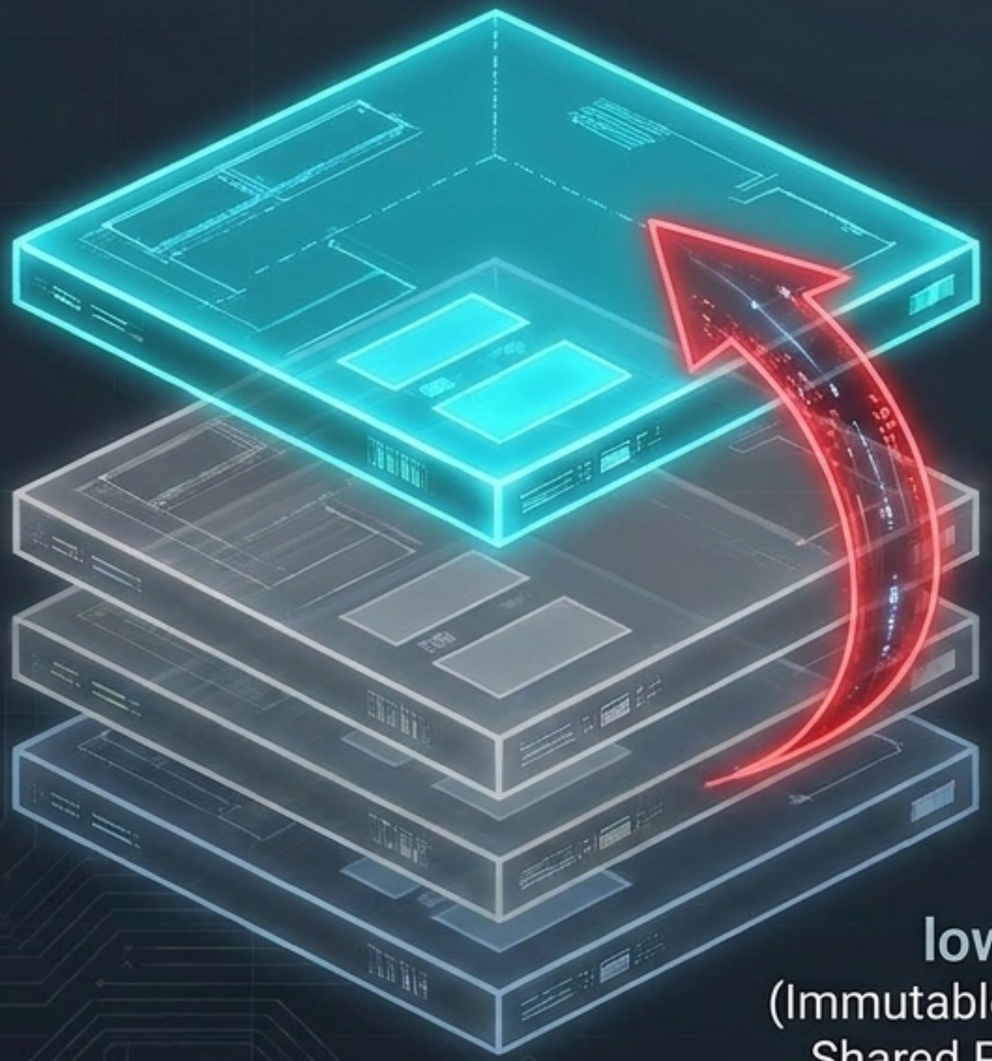



```
{
  "schemaVersion": 2,
  "mediaType": "application/vnd.oci.image.manifest.v1+json",
  "config": {
    "mediaType": "application/vnd.oci.image.config.v1+json",
    "digest": "sha256:c5c5220345714727389182760795213721116109836283243475691161093841",
    "size": 1284
  },
  "layers": [
    {
      "mediaType": "application/vnd.oci.image.layer.v1.tar+gzip",
      "digest": "sha256:a3ed95ca77b262715f8b00918483724239145991047731125155595938558893",
      "size": 38892
    },
    {
      "mediaType": "application/vnd.oci.image.layer.v1.tar+gzip",
      "digest": "sha256:8c022abf881433905906155824215551752235446867364136611030994160 27",
      "size": 21548
    },
    {
      "mediaType": "application/vnd.oci.image.layer.v1.tar+gzip",
      "digest": "sha256:e3c628894603884590608788790134926445986990207489023630389499193 1",
      "size": 74291
    },
    {
      "mediaType": "application/vnd.oci.image.layer.v1.tar+gzip",
      "digest": "sha256:b7d266434989253767441700395506884998815984746189577662964282 1916",
      "size": 42164
    }
  ]
}
```


An OCI Image is a standardized, content-addressable collection of SHA-256 compressed tarball layers, tied together by a JSON manifest.

Gemini Notebook

OverlayFS Architecture & Copy-on-Write Penalties



upperdir
(Ephemeral, Writable Container Layer)

Copy-on-Write (CoW) Penalty

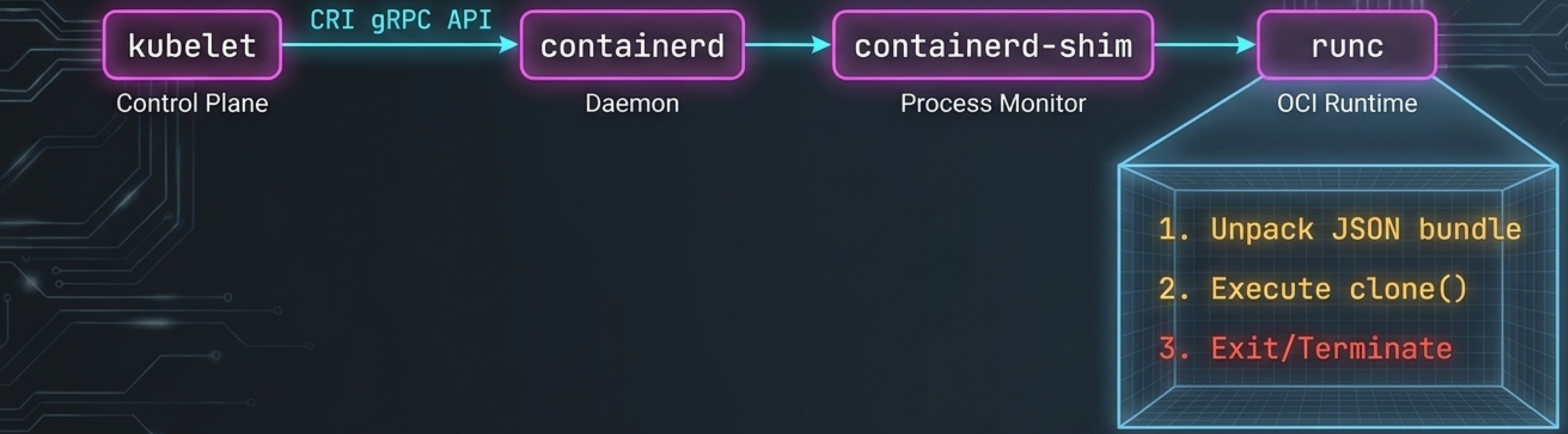
lowerdir
(Immutable OCI Layers / Shared Page-Cache)

The union mount merges directories logically.

Modifying an existing file triggers a Copy-on-Write penalty, forcing the kernel to duplicate the file into the upperdir before saving changes.

Gemini Notebook

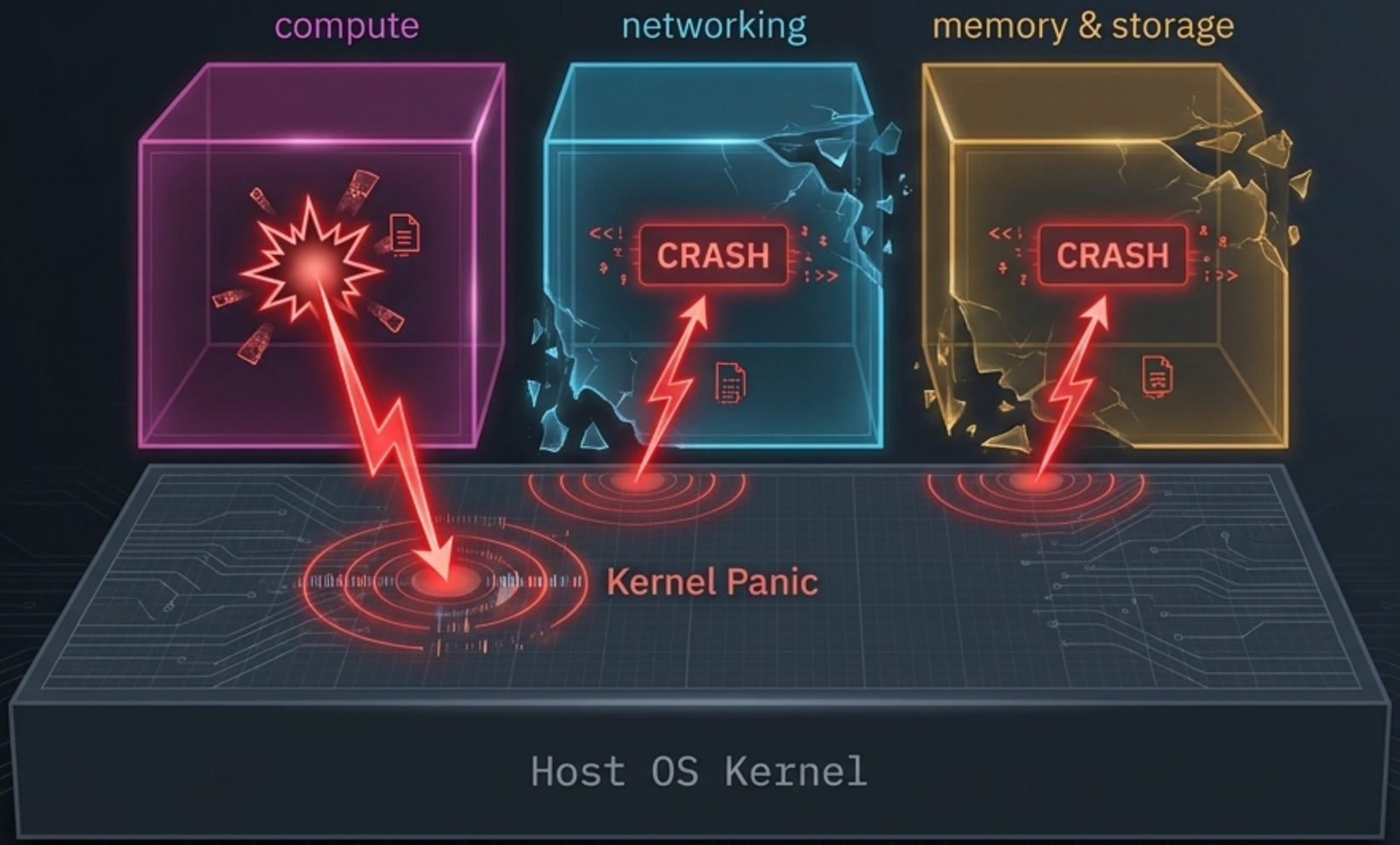
The CRI Execution Chain



The Container Runtime Interface (CRI) delegates execution. runc is the transient binary that physically unpacks the OCI bundle, configures the namespaces and cgroups, executes the entrypoint, and then terminates.

Gemini Notebook

The Shared Kernel Attack Surface



Standard containers prioritize density over deep isolation.

Because all tenants share the exact same kernel, a single kernel-level exploit compromises the entire host node and all neighboring workloads.

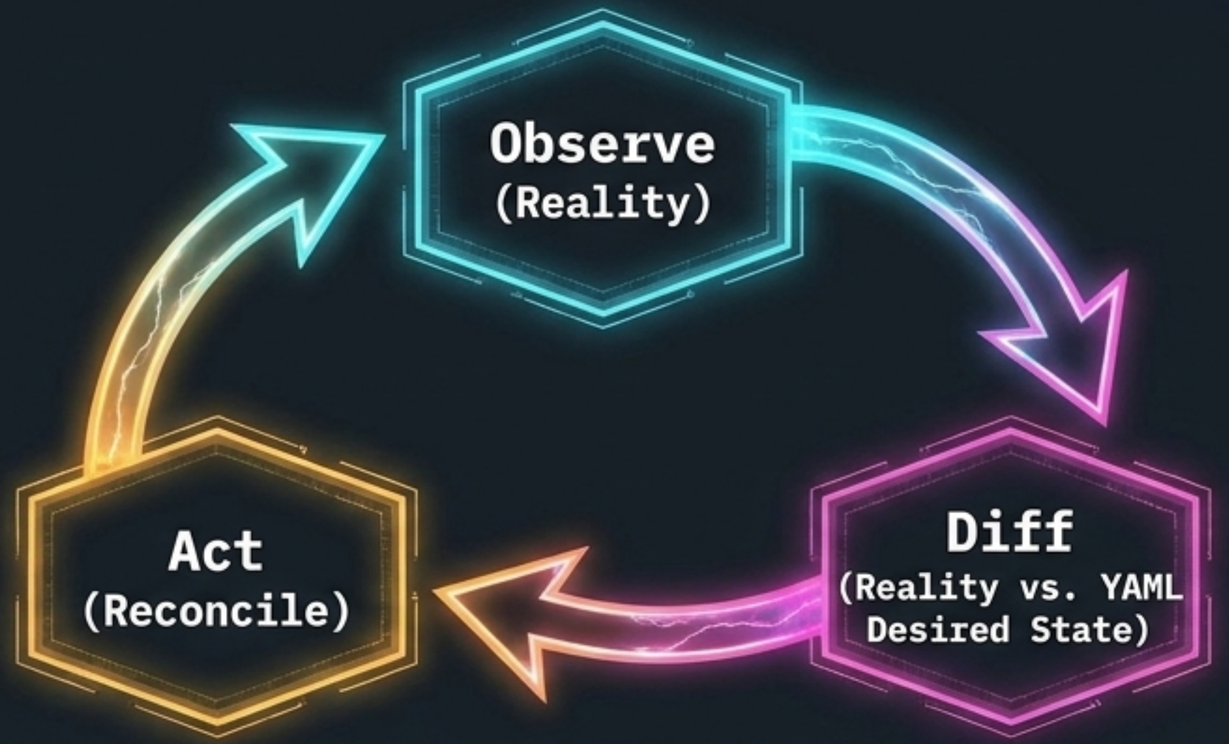
Gemini Notebook

Compute Selection: The Security Isolation Spectrum

Standard Containers (runc)	Virtual Machines (KVM)	MicroVMs (AWS Firecracker)	User-space Kernels (gVisor)
High Density	Low Density	High Density	Medium Density
Low Isolation (Shared Kernel)	High Isolation (Hardware Hypervisor)	High Isolation (Minimal device model)	Strong Isolation (ptrace/KVM syscall intercept)
Millisecond Boot	Slow Boot (Seconds)	Millisecond Boot	Moderate Performance Hit

Gemini Notebook

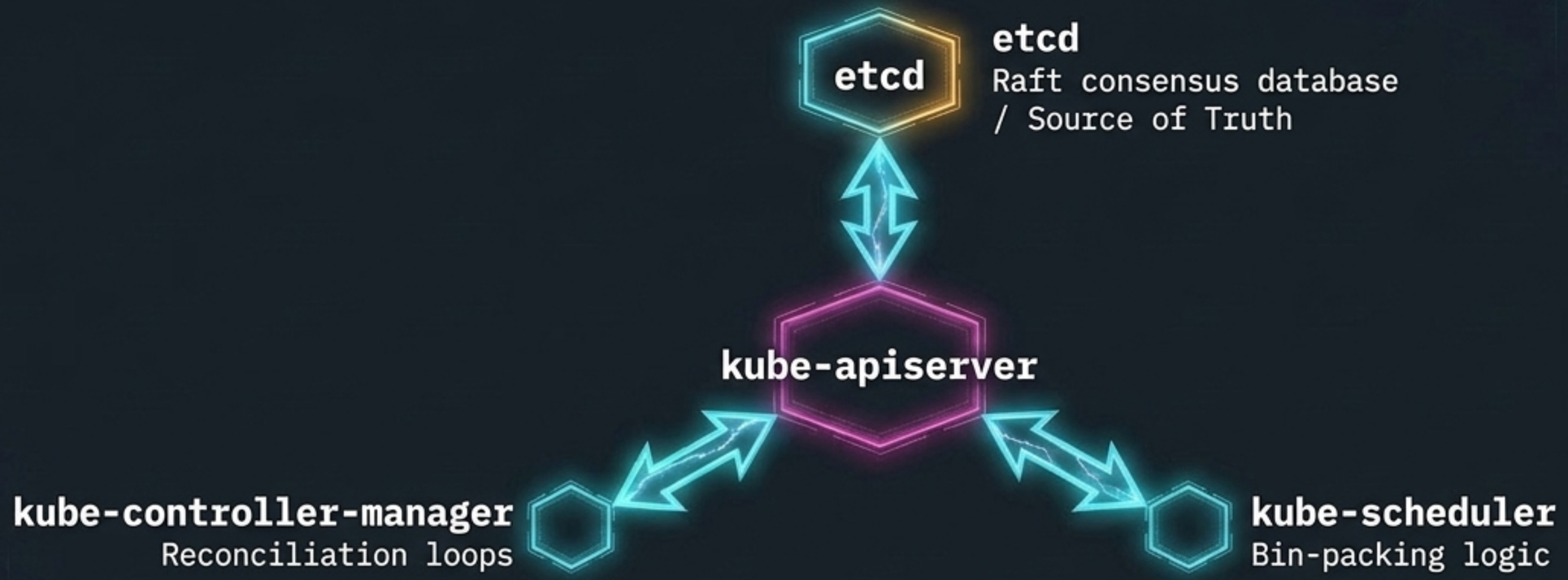
Level-Triggered Reconciliation



Kubernetes is not a workflow engine (A ➔ B ➔ C). It is a **declarative state reconciliation engine**. It continuously observes the actual state, diffs it against the desired state, and actively attempts to close the gap.

Gemini Notebook

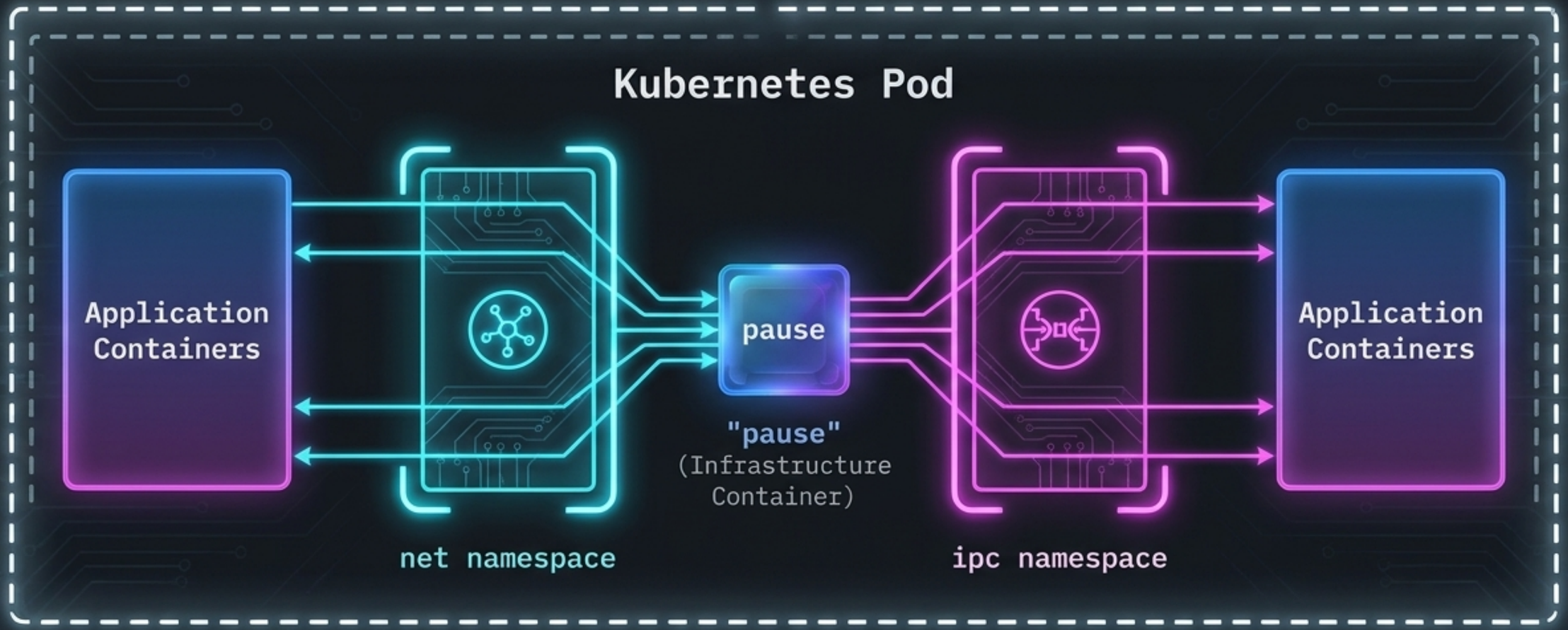
The Control Plane Brain



The **kube-apiserver** is the sole entry point. Components never communicate directly with each other; they strictly read from and write to the API server, with **etcd** serving as the cluster's single source of stateful truth.

Gemini Notebook

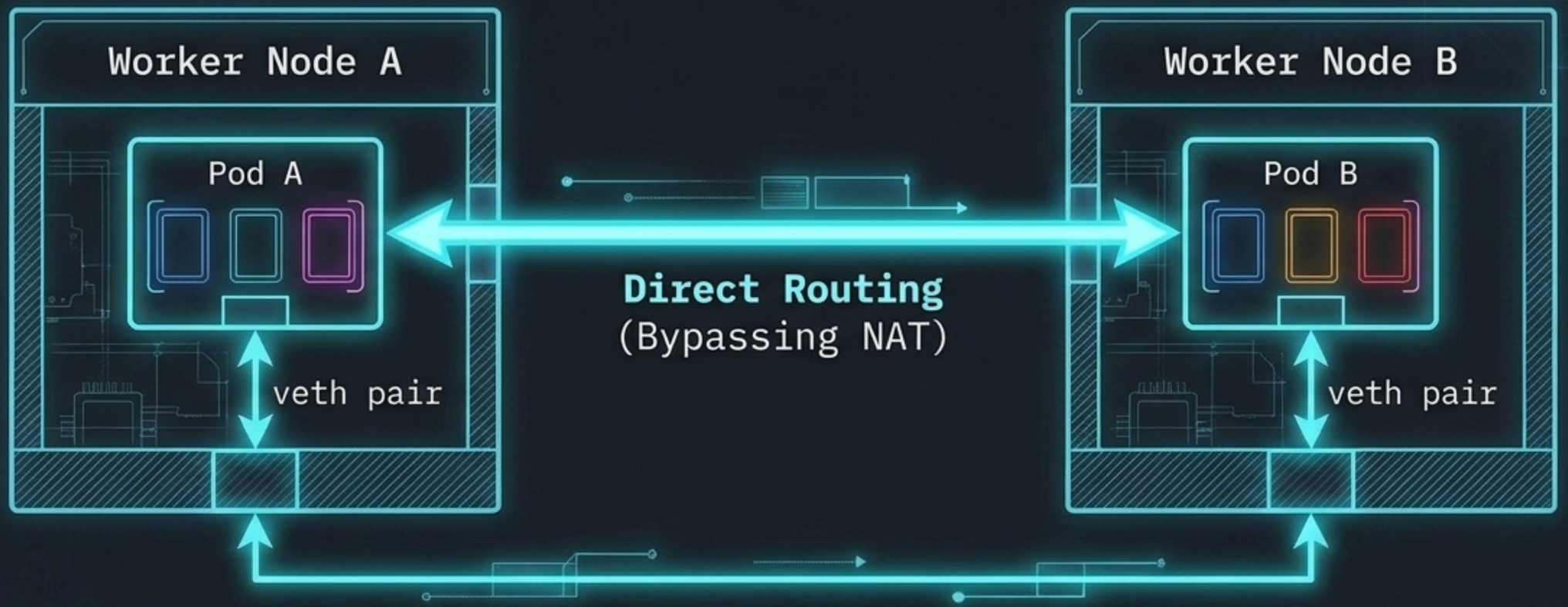
Deconstructing the Pod & The Pause Container



> Pods are tightly coupled container groups. The hidden "Pause" (infrastructure) container boots first to reserve and hold the network and IPC namespaces open, allowing application containers to restart or crash without losing the IP address.

Gemini Notebook

The Flat Pod IP Model & CNI



The **K8s networking mandate:** All Pods must communicate directly without NAT. The Container Network Interface (CNI) plugins (e.g., Cilium, VPC CNI) handle handle IP allocation and wire the veth pairs to the physical network.

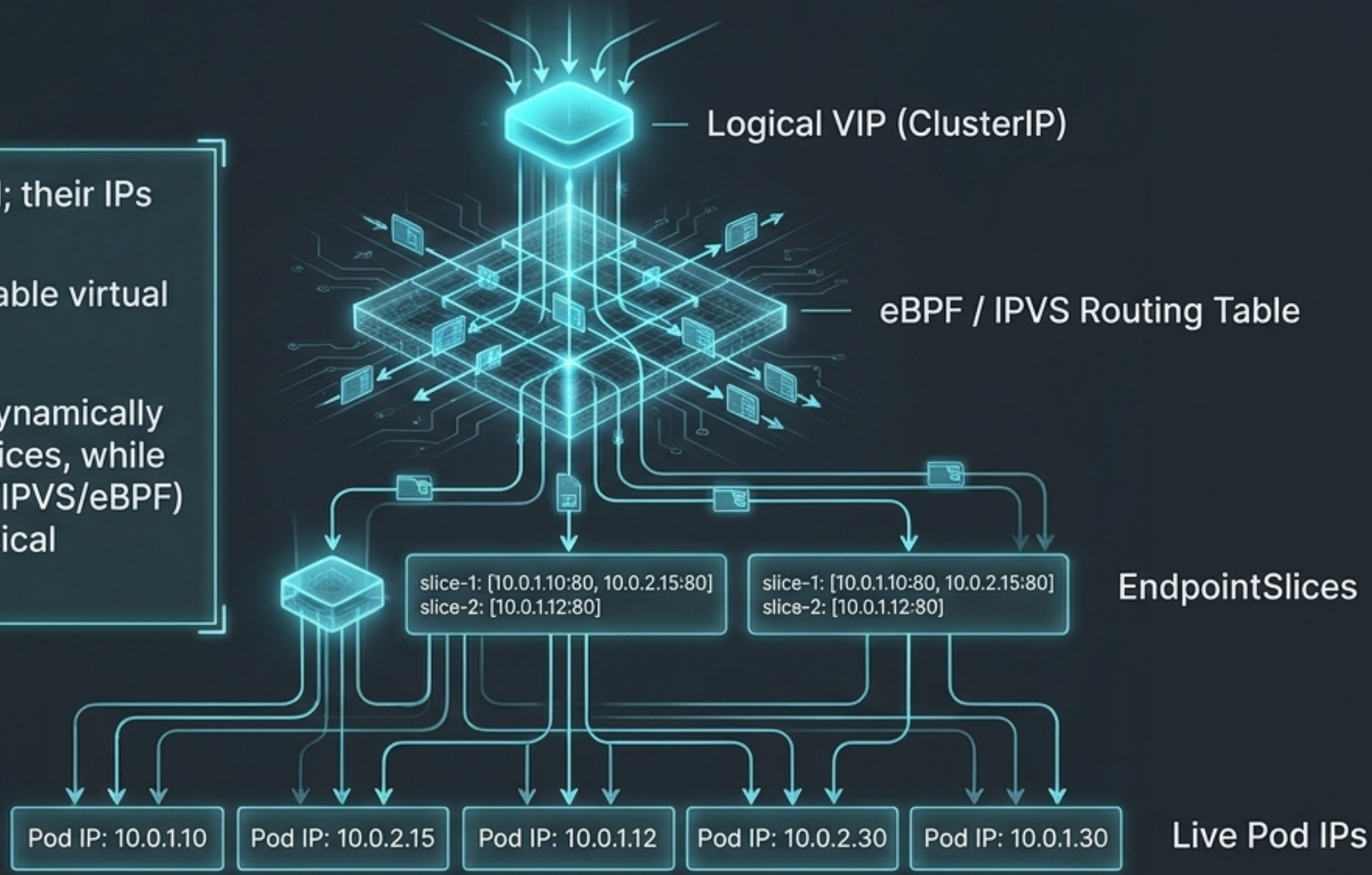
Gemini Notebook

Service Discovery & EndpointSlices

Pods are ephemeral; their IPs die with them.

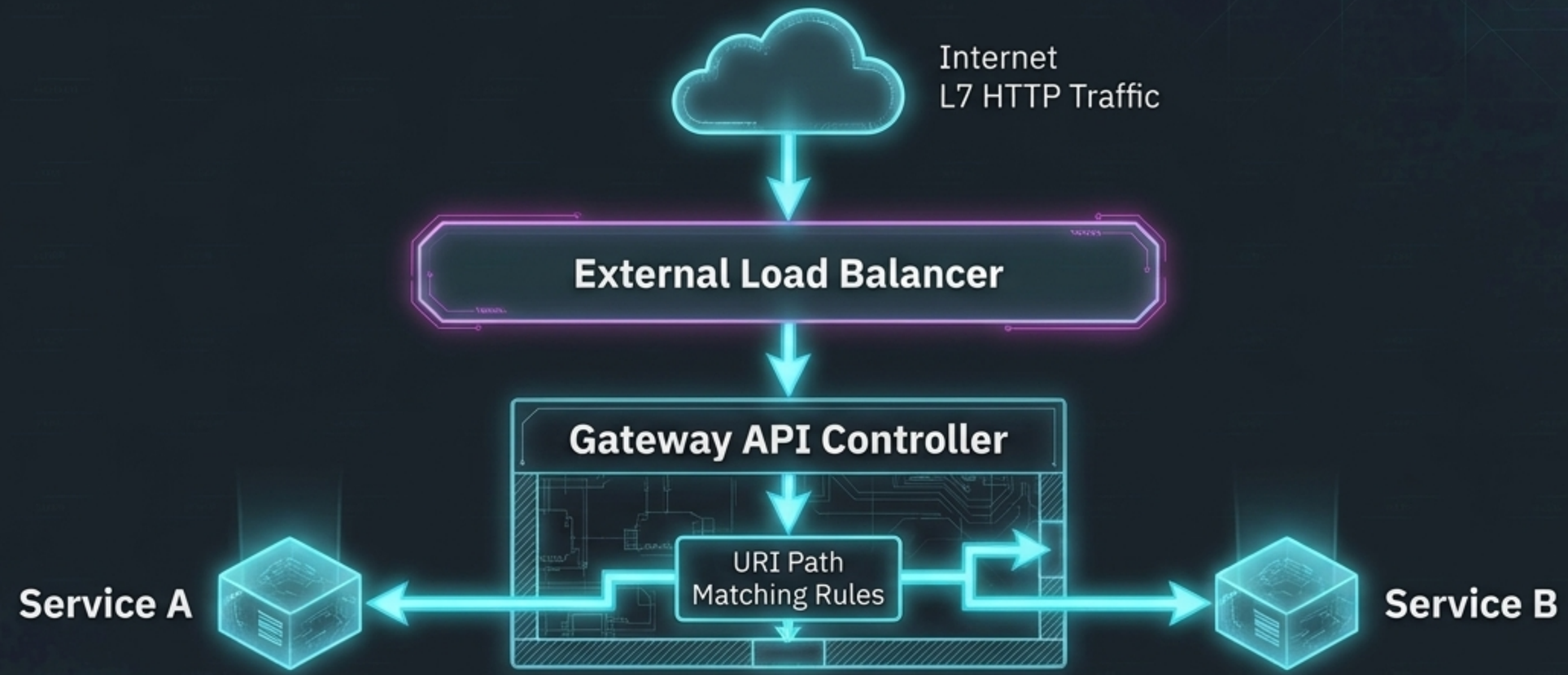
Services provide stable virtual IPs (ClusterIP).

The control plane dynamically updates EndpointSlices, while node-level proxies (IPVS/eBPF) implement the physical packet rewriting.



Gemini Notebook

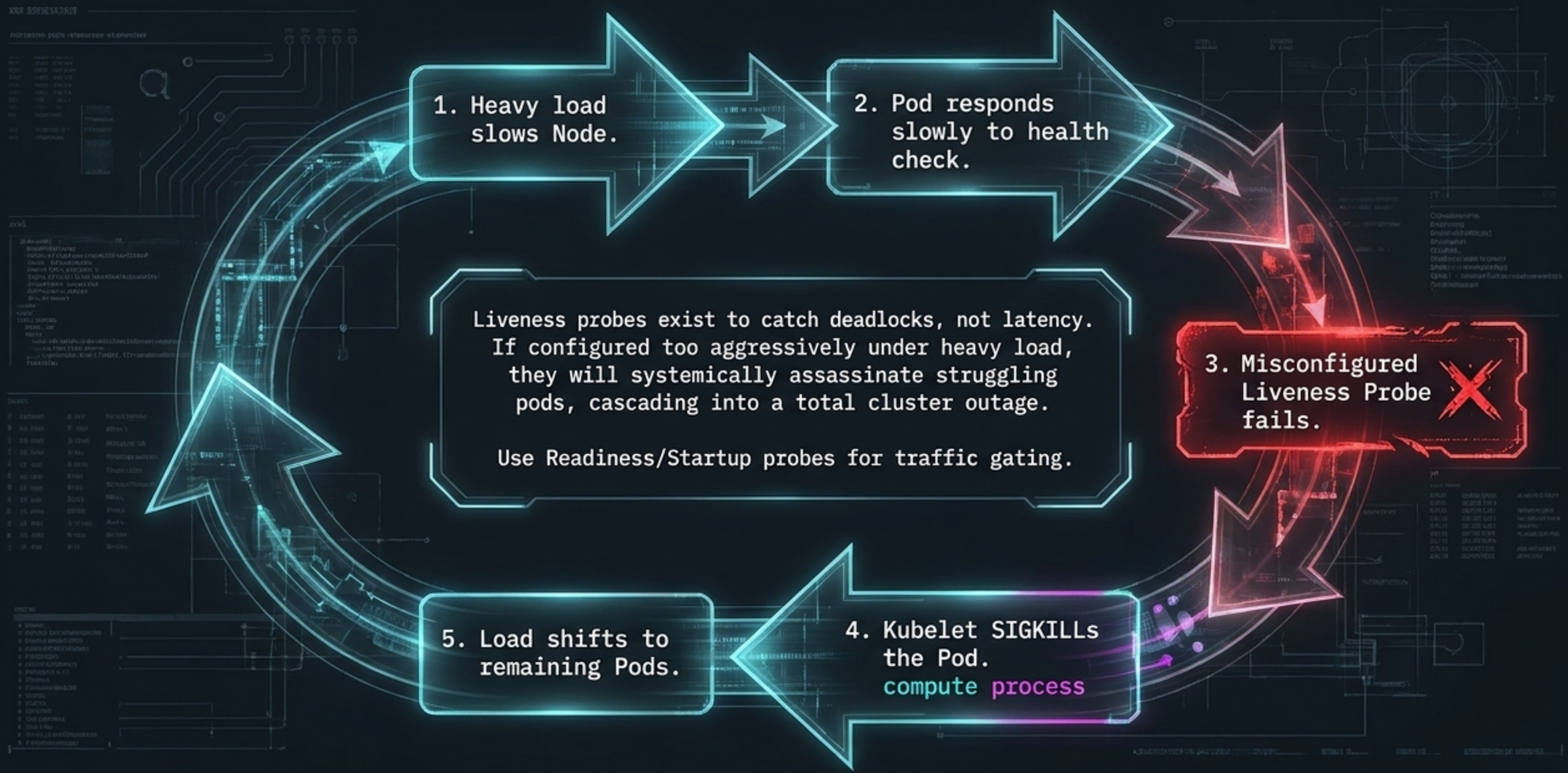
External Ingress & The Gateway API



The **modern evolution of K8s ingress**. The Gateway API separates infrastructure provisioning (Gateways) from application routing rules (HTTPRoutes), allowing platform engineers and developers to configure L7 routing independently.

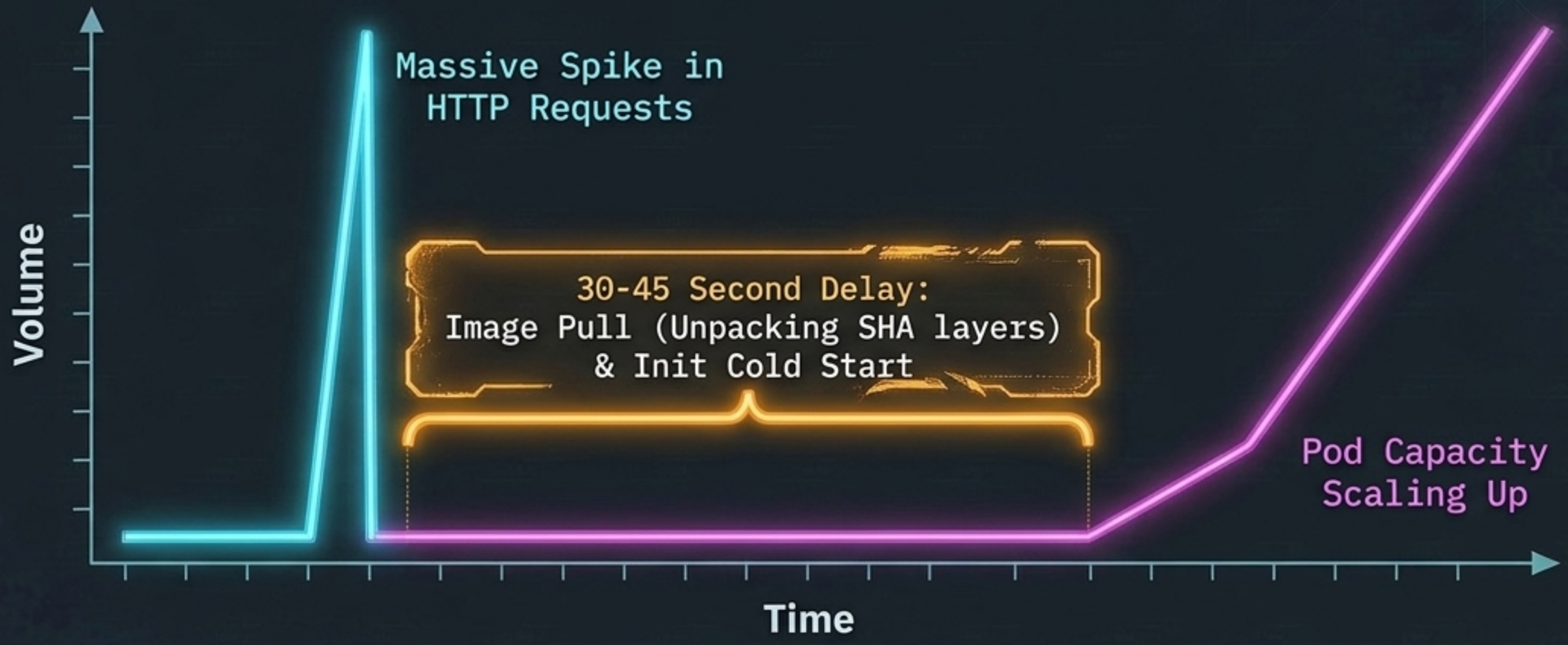
Gemini Notebook

Production Failure Modes: The Liveness Probe Cascade



Gemini Notebook

HPA Bursts & Cold Start Physics



Horizontal Pod Autoscalers (HPA) react to metrics, but the physics of scheduling, network image pulling, and application boot times dictate responsiveness. Massive layers ruin autoscaling agility.

Gemini Notebook

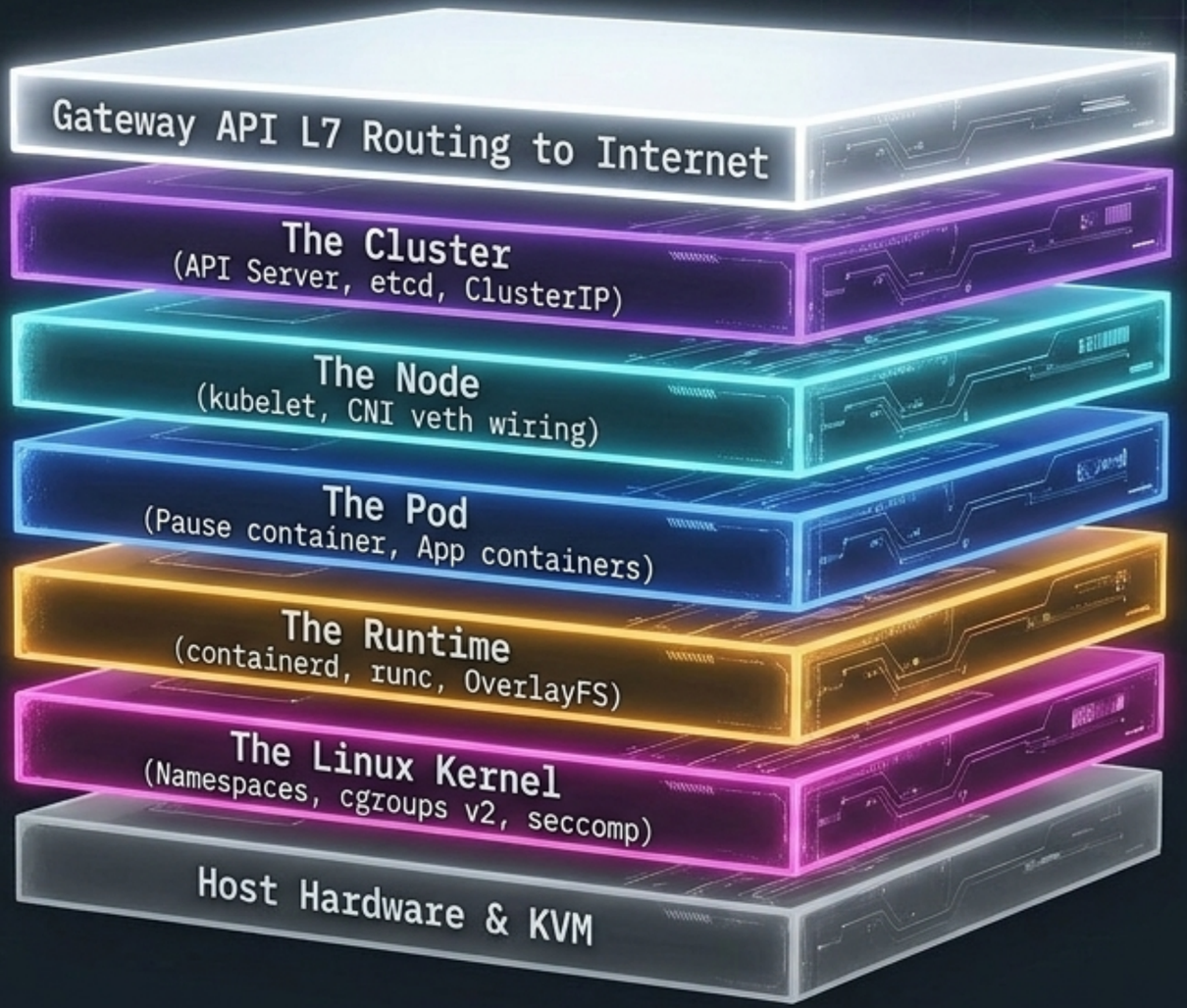
The Compute Selection Matrix

	Raw VMs	Managed Serverless (Cloud Run/ECS)	Kubernetes
Control	Maximum Kernel Control	Zero Ops Agility	Uniform API Across Clouds
Operational Overhead	Highest Maintenance Tax	Low	High Platform Complexity Tax
Ecosystem	Fragmented	Strict Vendor Lock-in	Massive Open Ecosystem

Gemini Notebook

Synthesis: The Cloud-Native Stack

The “magic” of the cloud does not exist. It is a meticulous, deeply layered orchestration of raw Linux system calls, virtualized network bridges, and declarative reconciliation loops.



Gemini Notebook

The X-Ray Diagnostic Checklist

[1] **Govern Resources**: Set **CPU Requests** for proportional weighting; avoid aggressive **CPU Limits** to prevent CFS throttling. Set **memory.max** accurately.

[2] **Optimize Images**: Shrink **OCI image layers** to eliminate HPA cold-start latency.

[3] **Isolate Intelligently**: Use **standard containers** for trusted multi-tenancy; default to **MicroVMs/gVisor** for untrusted code.

Gemini Notebook