



KubeCon



CloudNativeCon

Europe 2026

#KubeCon #CloudNativeCon

What happens in Kubernetes SIG Scalability?

Wojciech Tyczynski, Google



What do we do?

Define & Drive

Scalability definition & goals

Monitor & Measure

Performance of the system

Coordinate & Contribute

Performance improvements

Preserve & Protect

From scalability regressions

Consult & Coach

Community about scalability



KubeCon



CloudNativeCon

Europe 2026

Scalability Definition



What is Kubernetes scalability?



KubeCon



CloudNativeCon

Europe 2026

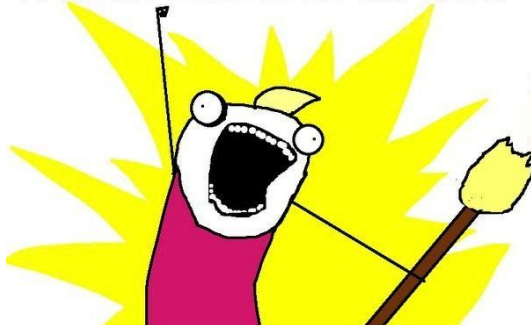
WHAT DO WE WANT?



SCALABLE CLUSTERS!



WHAT DOES IT MEAN?



What is Kubernetes scalability?



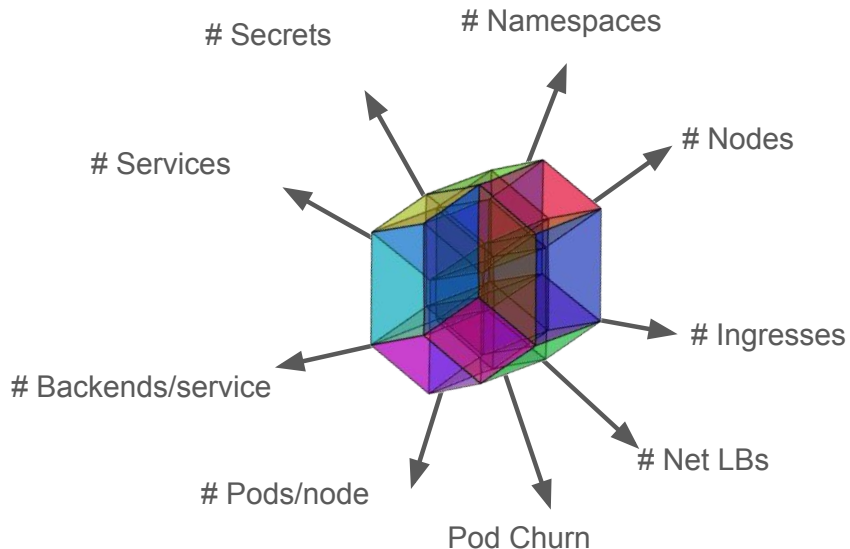
KubeCon



CloudNativeCon

Europe 2026

~~Scalability = # nodes~~



Scalability Envelope

Source of hypercube image:
<http://www.gregegan.net/APPLETS/29/29.html>

What is Kubernetes scalability?



KubeCon

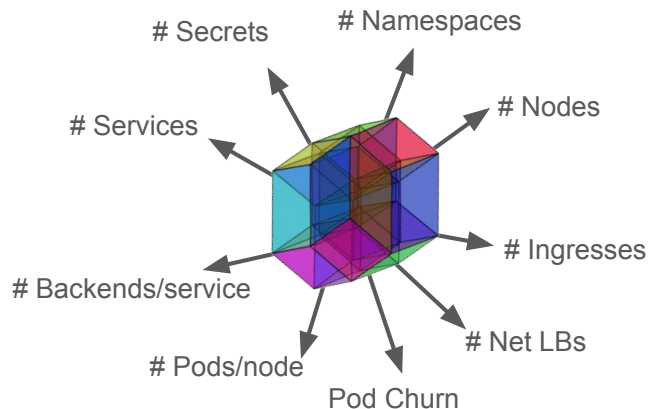


CloudNativeCon

Europe 2026

?

Scalability Envelope - a *safe zone*, within which your cluster is *happy*.



**ALL (scalability)
SLOs are satisfied**

Source of hypercube image:
<http://www.gregegan.net/APPLETS/29/29.html>

Kubernetes Scalability SLIs/SLOs



KubeCon



CloudNativeCon

Europe 2026

EXISTING INDICATORS

Core Measured SLIs:

- API Call Latency (Mutating & Read)
- Pod Startup Latency (End-to-end)
- [WIP] Network Programming Latency
- [WIP] DNS Programming & Latency

SLI VS SLO DEFINITION

SLI: Service Level Indicator. A quantitative measure of some aspect of the level of service provided.

SLO: Service Level Objective. A target value or range of values for a service level that is measured by an SLI.

CASE STUDY: API Call Latency Evolution

2015: INITIAL BLOG POST DEFINITION

SLO: "99% of all our API calls return in less than 1 second."

Simple, fleet-wide average. Lacked precision on object types or verbs.



2023: MATURE SYSTEM DEFINITION

SLI: Latency of processing mutating API calls for single objects for every (resource, verb) pair, measured as 99th percentile over last 5 minutes.

SLO: 99th percentile per cluster-day $\leq 1s$

Granular tracking per resource/verb. Excludes virtual/aggregated resources.

Kubernetes Scalability Limits



KubeCon



CloudNativeCon

Europe 2026

Scalability Envelope

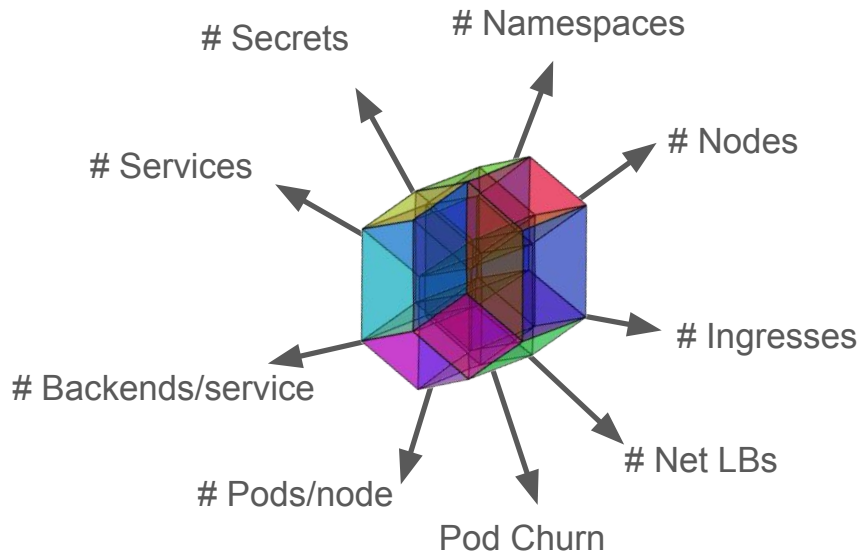
Precise definition of **scalability envelope** is ~impossible.

Fortunately, we have reasonable approximation:

- #Node $\leq$ 5,000
- #Pods $\leq$ 150,000
- #Service $\leq$ 10,000
- ...

Every usecase/scenario **requires its own validation**.

More at [sig-scalability/configs-and-limits/thresholds.md](https://sig-scalability/kubernetes/docs/configs-and-limits/thresholds.md)



Source of hypercube image:
<http://www.gregegan.net/APPLETS/29/29.html>



KubeCon



CloudNativeCon

Europe 2026

Scalability Testing



Scalability Testing Toolbox

ClusterLoader2

Declarative & Extensible

- **"Bring your own YAML"**: semi-declarative framework for complex test scenarios.
- **State-Driven**: Drives clusters to a desired state while verifying SLOs in real-time.
- **Deep Observability**: Integrated performance monitoring and bottleneck detection.

Designed for easy extensibility and deep performance insights.



Observability & Debugging

- Grafana: Comprehensive dashboards for real-time monitoring.
- Perfdash: Historical performance trend analysis and tracking.
- Profiling: Deep-dive into component-level CPU/memory bottlenecks.

Performance Guardrails

Systematic monitoring of SLIs (API Call Latency, Pod Startup) to ensure the scalability envelope is maintained across releases.

- **5,000 Node Tests:** Critical for detecting regressions early.

- Collaborative testing across major cloud providers via the community TestGrid.

- ## Specialized testing for features driven by relevant SIGs

[illegible]



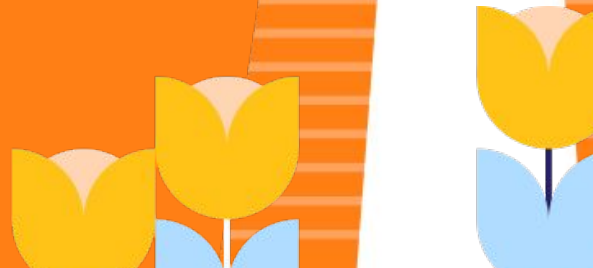
KubeCon



CloudNativeCon

Europe 2026

Scalability Improvements



The Two Pillars of Growth



KubeCon



CloudNativeCon

Europe 2026



Performance

1. Pushing the Limits

Expanding the hypercube of scalability by increasing absolute thresholds for nodes, namespaces, and pod density across the ecosystem.



Stability

2. Reliability at Scale

Ensuring the system remains healthy under load. Scalability is effectively reliability at scale, maintaining SLIs as the envelope grows.

Collaborative improvements driven with SIG API Machinery and SIG Scheduling.

KEP-3157: Streaming Watch-List



KubeCon



CloudNativeCon

Europe 2026

MECHANISM

From Chunked LIST to Streaming

Replaces traditional "LIST" requests with a continuous stream using the **SendInitialEvents=true** parameter.

- **Consistent RV:** Server may determine Resource Version from etcd first to prevent "time-travel".
- **Object Streaming:** Streams existing objects individually rather than as one giant chunk.
- **Bookmark Event:** Sends a bookmark containing the RV once the initial snapshot is complete.
- **Seamless Transition:** The informer continues the same WATCH request for ongoing updates.



SCALABILITY IMPACT

Protecting the API Server

Solves memory "explosions" that trigger OOM kills during heavy control plane load.

- **Predictable Memory:** Reduces footprint to ~2 MB per watcher regardless of object count.
- **Eliminates 5x Multiplier:** Removes the temporary allocation of 5x response size in API server memory.
- **Reduced etcd Load:** Leverages watch-cache streaming to decrease direct etcd pressure.
- **Stale Read Fix:** Ensures data consistency for informers starting from the cache.

KEP-5116: Streaming LIST Encoding



KubeCon



CloudNativeCon

Europe 2026

MECHANISM

Chunked Encoding for Collections

Replaces monolithic response marshalling with a per-object streaming encoder for **JSON** and **Protobuf**.

- **Iterative Marshalling:** Processes individual items within the "Items" array instead of allocating a single contiguous block.
- **Strict Validation:** Validates Go struct tags to ensure byte-for-byte compatibility with standard encoders.
- **Gzip Optimization:** Rewritten compression buffers small writes to maintain threshold conformance while supporting chunks.
- **Fallbacks:** Automatically reverts to standard encoders if custom tag validation fails for built-in or custom types.



SCALABILITY IMPACT

Predictable Memory & Stability

Directly addresses API server memory exhaustion caused by large resource (e.g., CRD) LIST requests.

- **OOM Prevention:** Eliminates GB-scale spikes by keeping only the active object (max 1MB) in encoding memory.
- **Cost Efficiency:** Reduces need for over-provisioning RAM to handle concurrent naive controller requests.
- **Performance Neutral:** Trades slight CPU/allocation increases for massive reductions in peak memory usage.
- **Reliability Guard:** Introduces release-blocking scalability tests to detect and prevent future memory regressions.

KEP-4988: Snapshottable APIServer Cache



KubeCon



CloudNativeCon

Europe 2026

MECHANISM

B-Tree Snapshots for Historical LISTs

Enables serving historical **LIST** requests (pagination/RV) directly from the watch-cache instead of etcd.

- **Lazy Copying:** Uses B-tree Clone() for efficient, pointer-based point-in-time snapshots.
- **Aligned Compaction:** Snapshots are truncated following etcd's compaction cycle to maintain conformance.
- **Auto-Detection:** Periodic hashing of resources between etcd and cache to detect bugs.
- **Safety Fallback:** Automatically reverts to etcd if a cache inconsistency is detected.



SCALABILITY IMPACT

Taming Memory & Load

Significantly reduces the stability risks associated with large-resource LIST requests.

- **Predictable RAM:** Eliminates volatile memory pressure from direct etcd data transfers.
- **Low Overhead:** Testing shows only ~1.3% total memory increase and 0.2% allocation growth.
- **Enhanced Throttling:** Makes APF more effective by shifting load to memory-stable cache reads.
- **Etcd Relief:** Dramatically reduces heavy read load on etcd for common paginated workloads.

KEP-5647: Stale Controller Handling



KubeCon



CloudNativeCon

Europe 2026

MECHANISM

Read-After-Write Consistency Logic

Introduces a mechanism for controllers to ensure their local cache has observed previous writes before re-reconciling.

- **Informer Bookmarks:** Adds a **BookmarkFunc** to informers to track resourceVersion (RV) lifecycle accurately.
- **RV Mapping:** Stores a mapping of reconciled objects to a tuple of expected RVs for key dependencies (e.g., Pods).
- **Smart Requeue:** Automatically skips and requeues reconciles with exponential backoff if the cache is stale.
- **Opt-in Implementation:** Targeted at high-churn controllers like DaemonSet and StatefulSet first.



SCALABILITY IMPACT

Mitigating Eventual Consistency Risks

Protects cluster stability by preventing "hot loops" and duplicate work during heavy control plane lag.

- **OOM Prevention:** Eliminates spurious reconciles that can trigger memory "explosions" under load.
- **Reliable Scalability:** Ensures controllers don't duplicate objects (like extra Pods) when the watch stream falls behind.
- **Reduced API Stress:** Prevents unnecessary write operations based on outdated cache state.
- **Standardized Metrics:** (Alpha Non-Goal) Future monitoring will expose quantified "cache staleness" seconds for operators.

KEP-5598: Opportunistic Batching



KubeCon



CloudNativeCon

Europe 2026

MECHANISM

Pod Signatures & Cached Results

Uses pod signatures to identify "identical" scheduling requirements and reuse previous work.

- **PodSignature:** Fragments from all plugins are marshaled into a unified byte slice signature.
- **GetNodeHint:** Validates cache freshness and ensures the "last node is full" for 1-pod-per-node efficiency.
- **Fast Path:** Hinted nodes bypass full iteration and scoring phases entirely if they pass filters.
- **StoreScheduleResults:** Records sorted feasible nodes from "canonical" pods for future reuse.



SCALABILITY IMPACT

Breaking the $O(N*M)$ Barrier

Dramatically reduces scheduling cost for large jobs on large clusters with simple constraints.

- **$O(1)$ Path:** Transitions from per-node computation to single map lookups for compatible pod batches.
- **Reduced Latency:** Pre-calculating signatures in the queue removes overhead from the time-sensitive scheduling path.
- **Batch/ML Optimized:** Specifically targets the "1-pod-per-node" reality of modern high-scale AI workloads.
- **Performance Baseline:** Validated via dedicated scheduler_perf scenarios like HostPortConflict and ResourceSaturation.

Issue-134375: Resource Size Dimension



KubeCon



CloudNativeCon

Europe 2026

THE PROBLEM

Stale Assumptions vs. Reality

Scalability tests have historically focused on hyper-optimized 1-3 KB Pod objects.

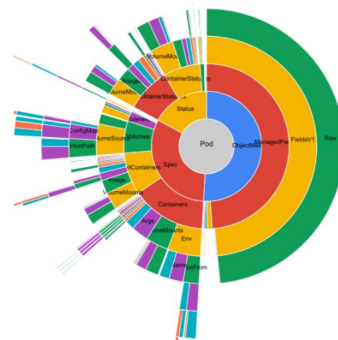
- **Resource Bloat:** Modern AI/ML workloads utilize sidecars and complex features.
- **User Reality:** Real users often use Pods that can achieve up to 100 KB of serialized representation.
- **The Gap:** Unblocking larger AI workloads requires addressing the limits that users are hitting related to large amount of data.

SCALABILITY IMPACT

When Limits Break, They Fail Hard

Surpassing untested limits leads to immediate, unrecoverable failures.

- **Catastrophic Downtime:** Kubernetes "breaks" rather than "bending" under resource size load.
- **Failure Loops:** Load causes lock contention and data staleness, triggering reset loops.



How to get involved?



KubeCon



CloudNativeCon

Europe 2026

Where to find us?

- Home page: [README](#)
- Public Meetings: Thursdays 17.30 UTC (bi-weekly)
- Slack channel: [#sig-scalability](#)
- Mailing List: [#kubernetes-sig-scale](#)



KubeCon



CloudNativeCon

Europe 2026



Connecting Your Laptop to Present



KubeCon



CloudNativeCon

Europe 2026

If you plan to use speaker notes, please follow the steps below to ensure your laptop is set up correctly. This will allow you to view your notes while the audience sees only your slides.

Before Connecting

- Close all unnecessary apps & tabs
- Turn off notifications
- Disable Night Mode / TrueTone
- Don't enter full-screen presentation mode yet

At the podium

- Plug in the HDMI cable
- If needed, use your USB-C to HDMI adapter
- On Mac, click "Allow" if prompted

USE "EXTENDED DISPLAY" MODE (NOT MIRRORED)

Using MAC

- Go to System Settings > Displays
- Click the "Arrangement" tab
- Make sure "Mirror Displays" is unchecked

Using PC

- Press $\square$ Win + P, then select "Extend"
- Or: Right-click on the Desktop > Display Settings
- Under Multiple Displays, select "Extend these displays"

Connecting Your Laptop to Present



KubeCon



CloudNativeCon

Europe 2026

When Using PowerPoint

- Start slideshow: ⌘Cmd + Enter (Mac) or F5 (PC)
- Or: Click Slide Show > Play from Start
- If notes show on the wrong screen, click “Swap Displays”
- <https://www.youtube.com/watch?v=gQ3D4m-5pww>

Google Slides

- Use Google Chrome as your browser
- Click the dropdown next to ‘Slideshow’ > select Presentation Display Options
- Check “Presenter View” and “Full Screen”
- If prompted, select "Allow"
- Choose the external display (may have a strange name like PanasonicXYZ)
- Click Start Slideshow
- <https://www.youtube.com/watch?v=-GT7WCvPcys>