

Accelerating Broadcast-Quality Video at Scale with VPU's on Akamai Cloud



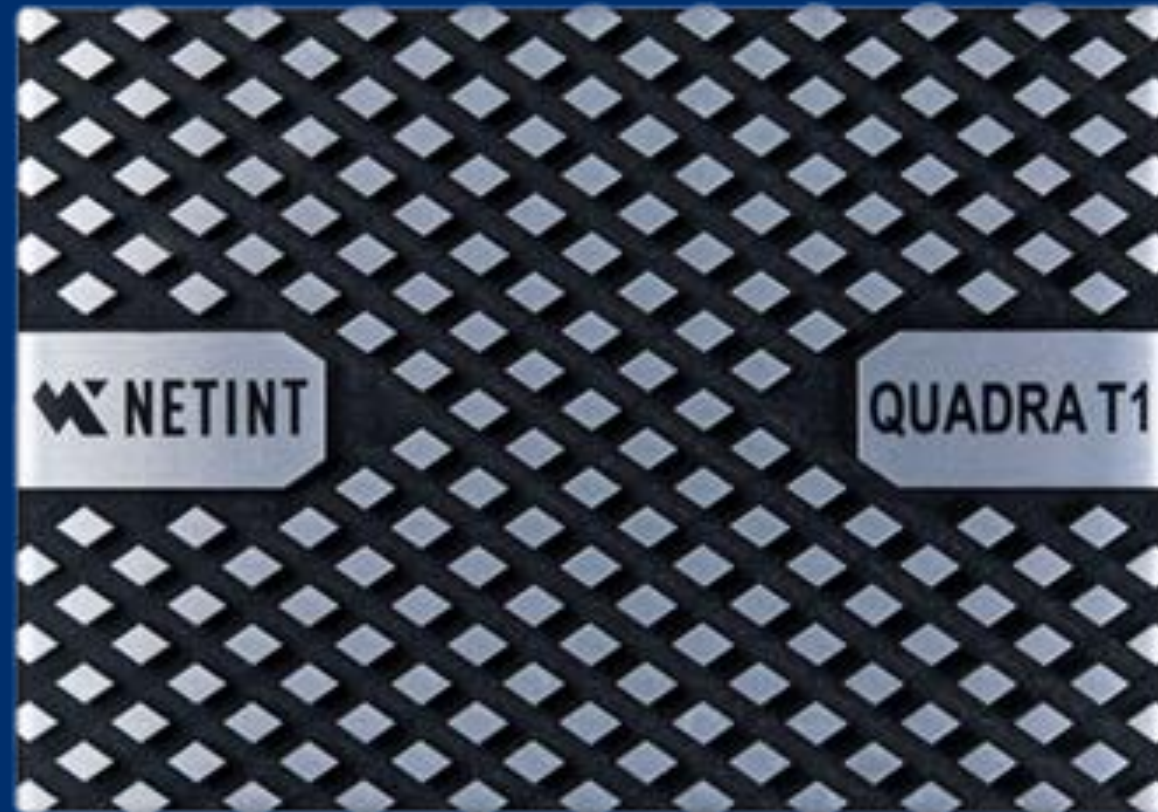
Anders Näsman

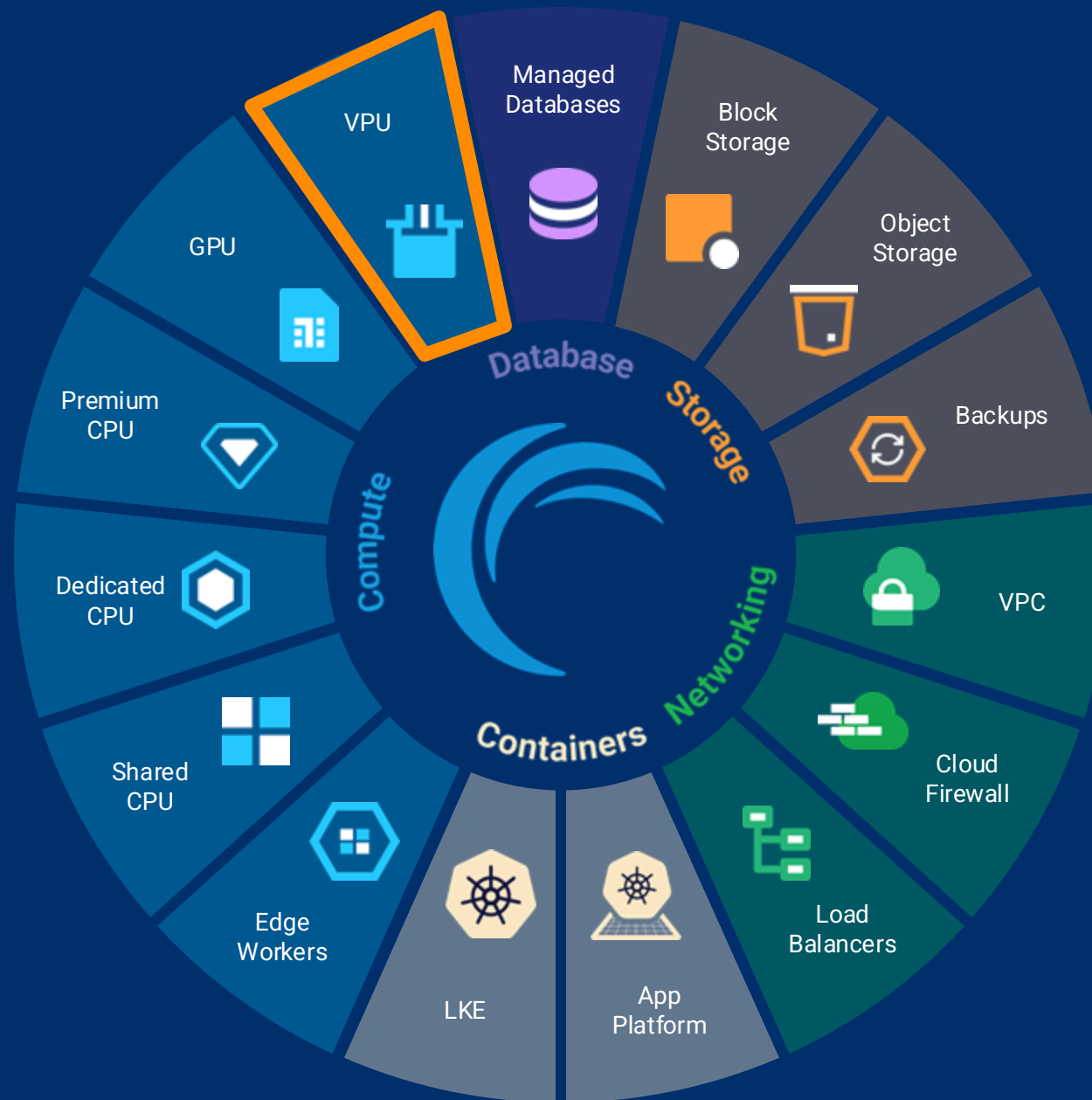
Principal Solutions Engineer, Akamai
anasman@akamai.com



Dominique Vosters

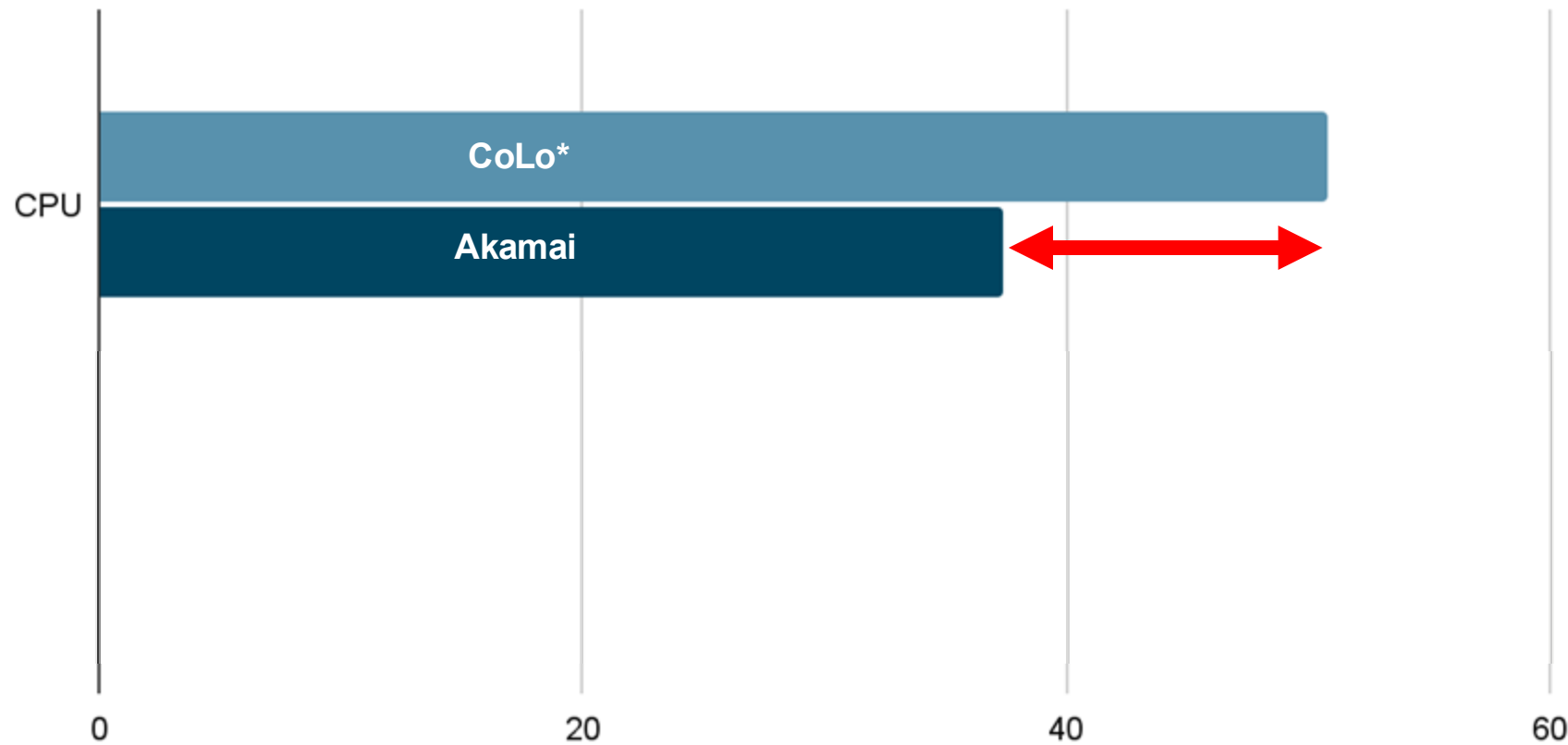
Sales Director, Scalstrm
dominique@scalstrm.com







Average kWh/day

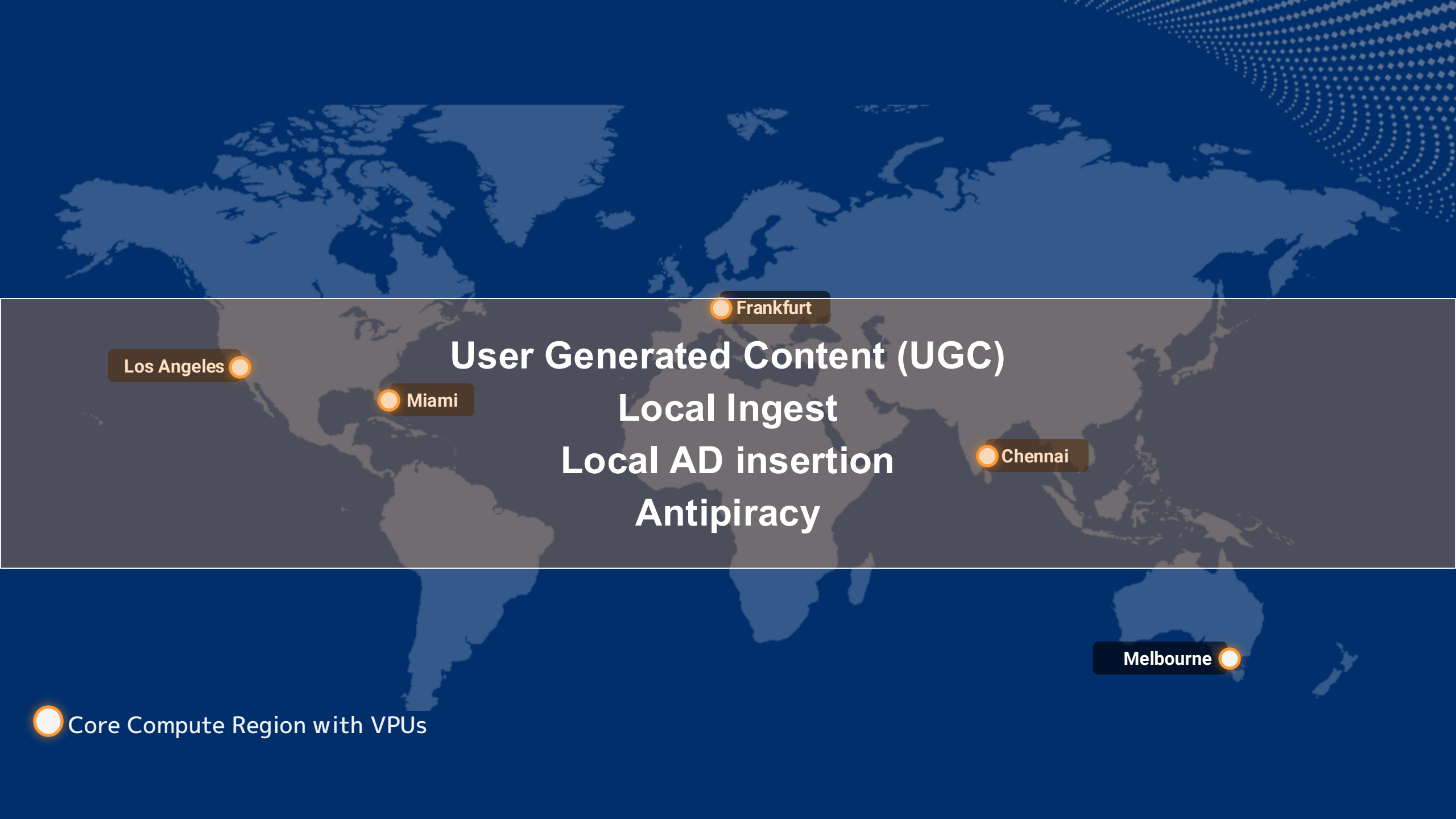


* CoLo CPU: 3 x Supermicro SYS-111C-NR without VPU/GPU,
CoLo VPU: Supermicro SYS-111C-NR + 1x Quadra T1U



● Core Compute Region

● Core Compute Region with VPUs



User Generated Content (UGC)
Local Ingest
Local AD insertion
Antipiracy

Los Angeles

Miami

Frankfurt

Chennai

Melbourne

Core Compute Region with VPUs

1,000

TBPS CAPACITY

4,100+

EDGE PoPs

1,200+

NETWORKS

26

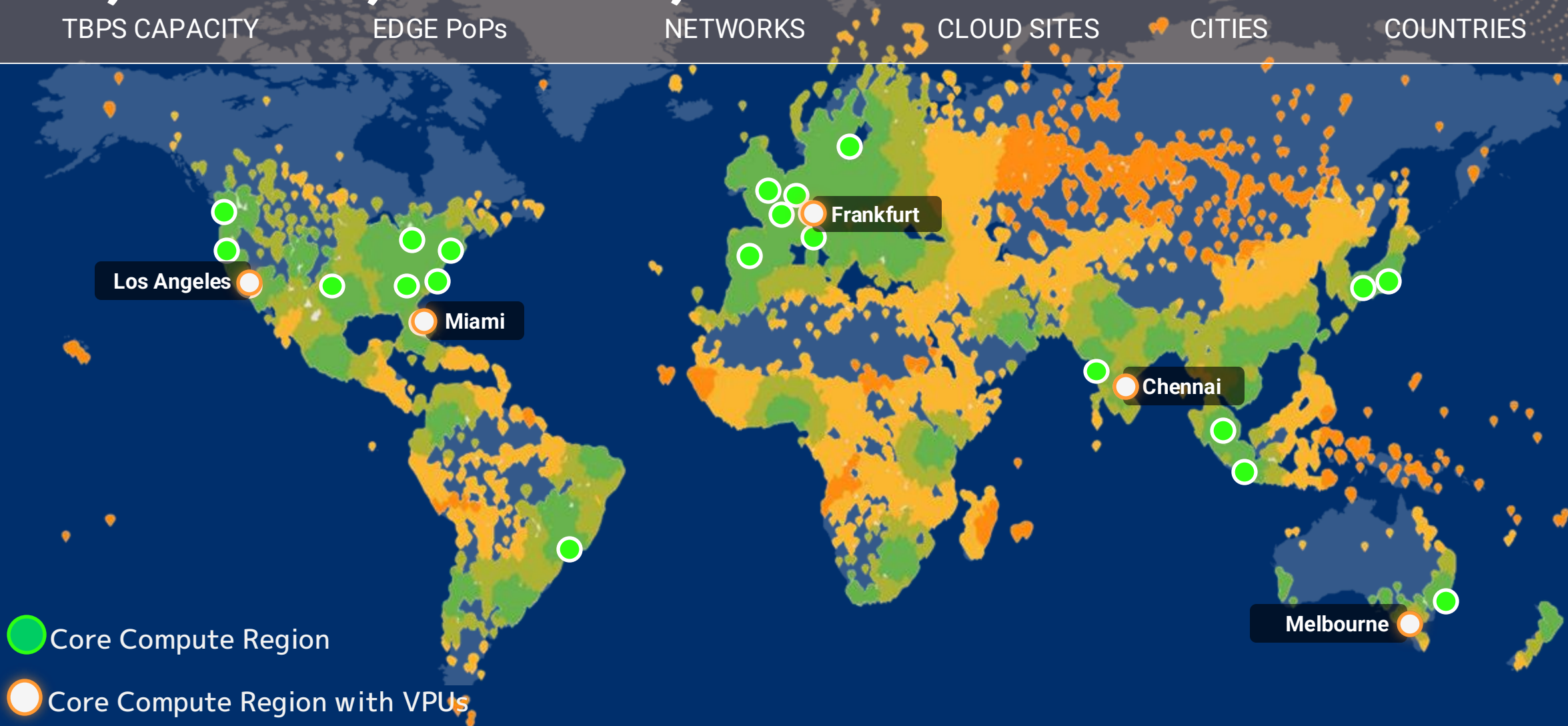
CLOUD SITES

750+

CITIES

130

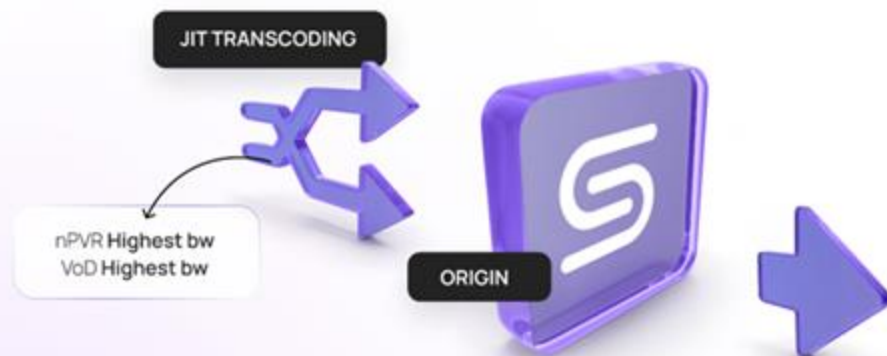
COUNTRIES





JUST-IN-TIME TRANSCODING

Unlock the future of video streaming with Scalstrm's VOD Just-In-Time Transcoding. Eliminate the need for excessive pre-transcoding and storage of multiple profiles. With on-the-fly video optimization, Scalstrm enables cost-effective, efficient, and sustainable video delivery tailored to modern demands.



Scalstrm's innovative approach saves 50-60% of storage space by keeping all profiles for the first 2-3 days when content is most popular. After this period, only the highest-quality profile remains stored, while lower profiles are transcoded on demand for long-tail content.

- 
- ✓ 0% Pre-Transcoding of entire libraries
 - ✓ Storage Optimization
 - ✓ Seamless Integration
 - ✓ Flexible deployment
 - ✓ Dynamic Performance



HW acceleration through NETINT cards



LIVE TRANSCODING

Dramatically Reduce Cost

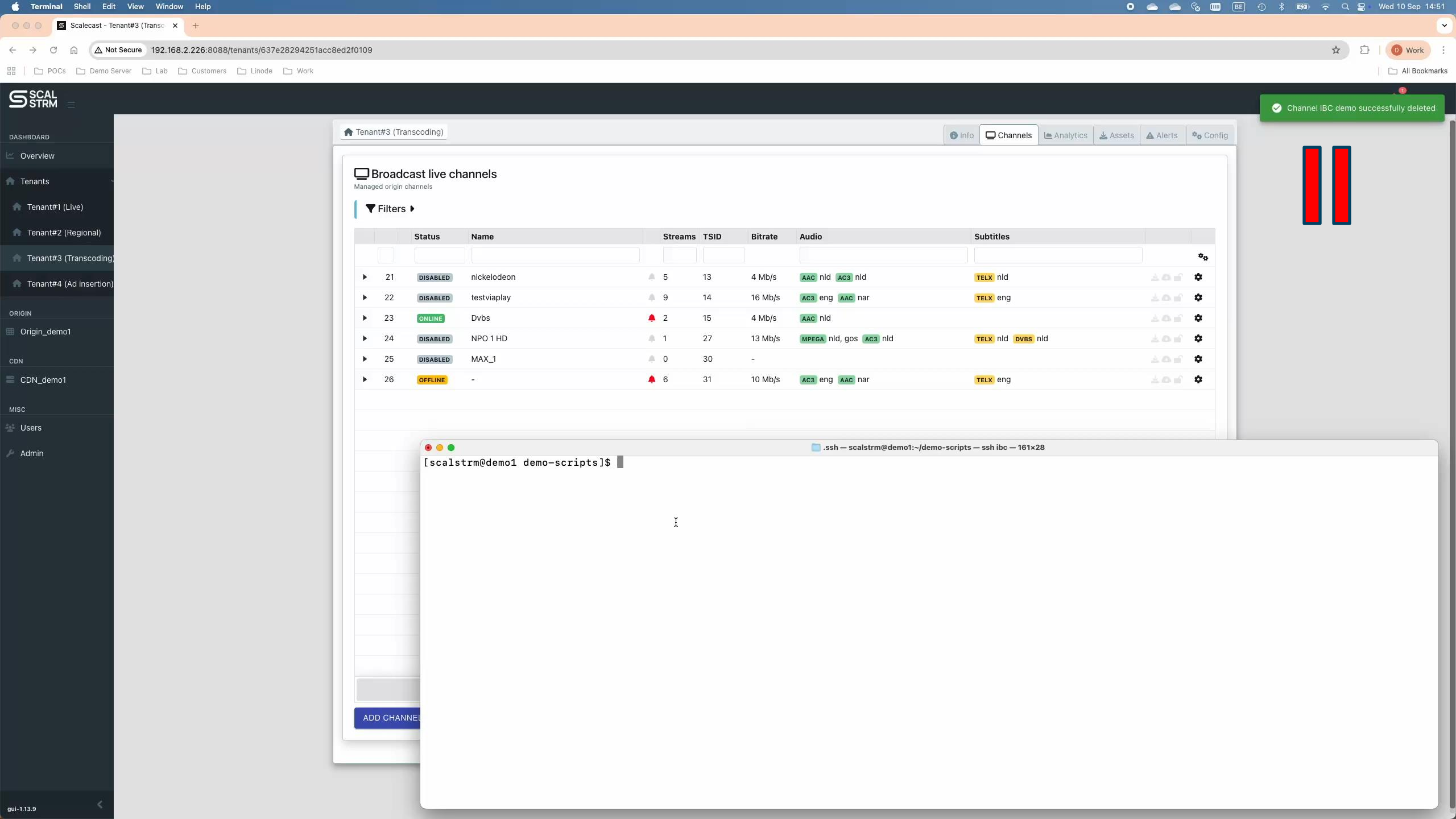
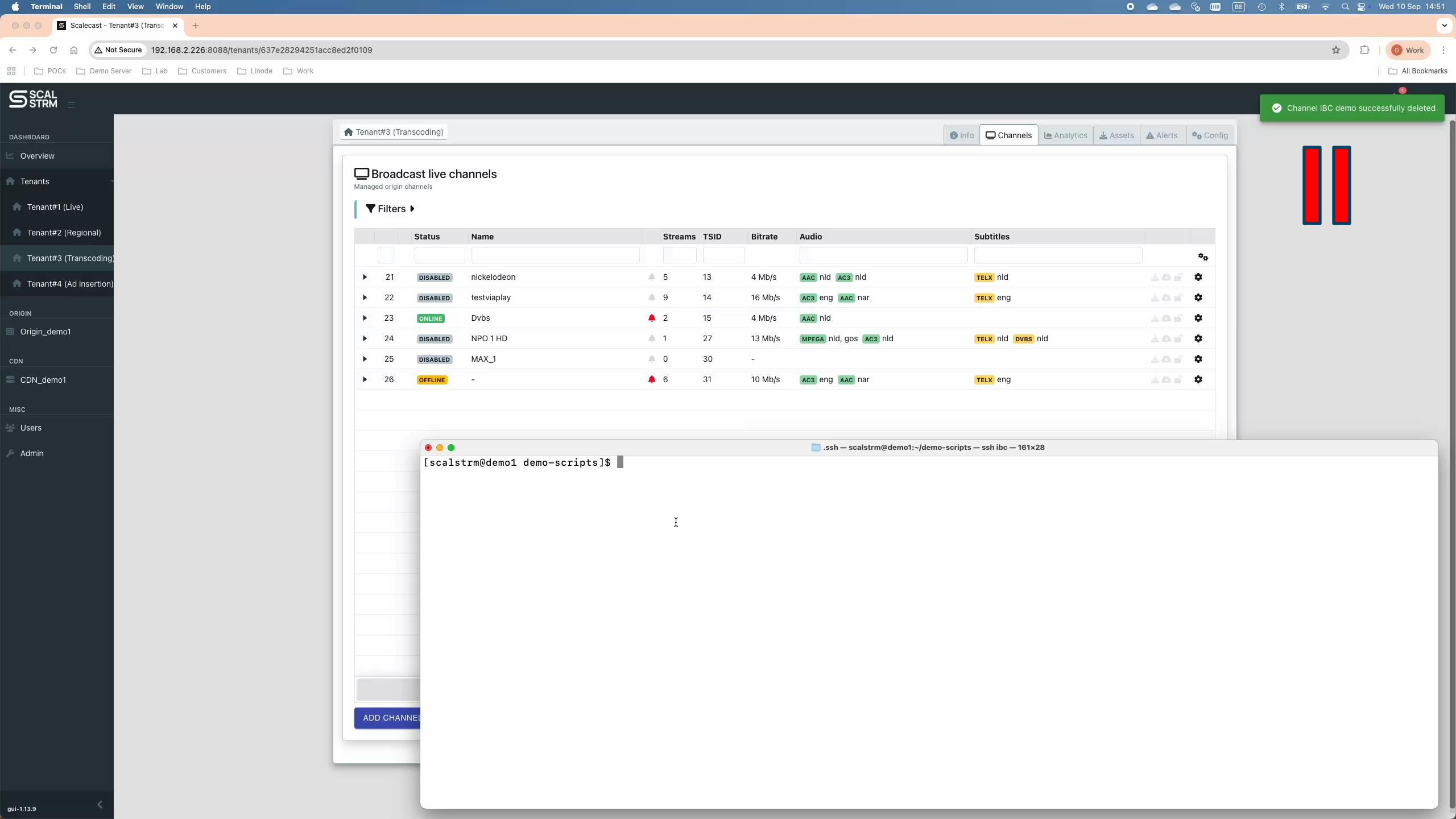
Transcode up to 4K HEVC using the latest technology e.g. VPU with highly flexible orchestration for event-based channels.



Ensure reliable, high-quality live streaming with a platform designed to simplify operations and maximize performance.

- ✓ Single Video Processing Platform:
Integrates seamlessly into the Scalstrm Origin and CDN platform
- ✓ Easy to use scheduling tool / GUI
build into our interface for 24 x 7 and event channels
- ✓ Latest Transcoding Tech
CPU / VPU / GPU Support - Multi-card scheduling
- ✓ Event Channels via Auto-scaling in clouds
stateless so starts up automatically from schedule in a few seconds
- ✓ Comprehensive Monitoring, built-in trouble-shooting tools

Scalstrm's Single Video Processing Platform
Seamlessly integrated, auto-scaling, and built
for 24/7 live transcoding.



DIFFERENTIATORS

Performance

Highly optimized components - scale independently
800 ABR channels, 200 Gbps egress per server

Flexibility

Modern elastic micro component architecture.
Deploy on premise, private or public cloud, multi clouds, or any combination.

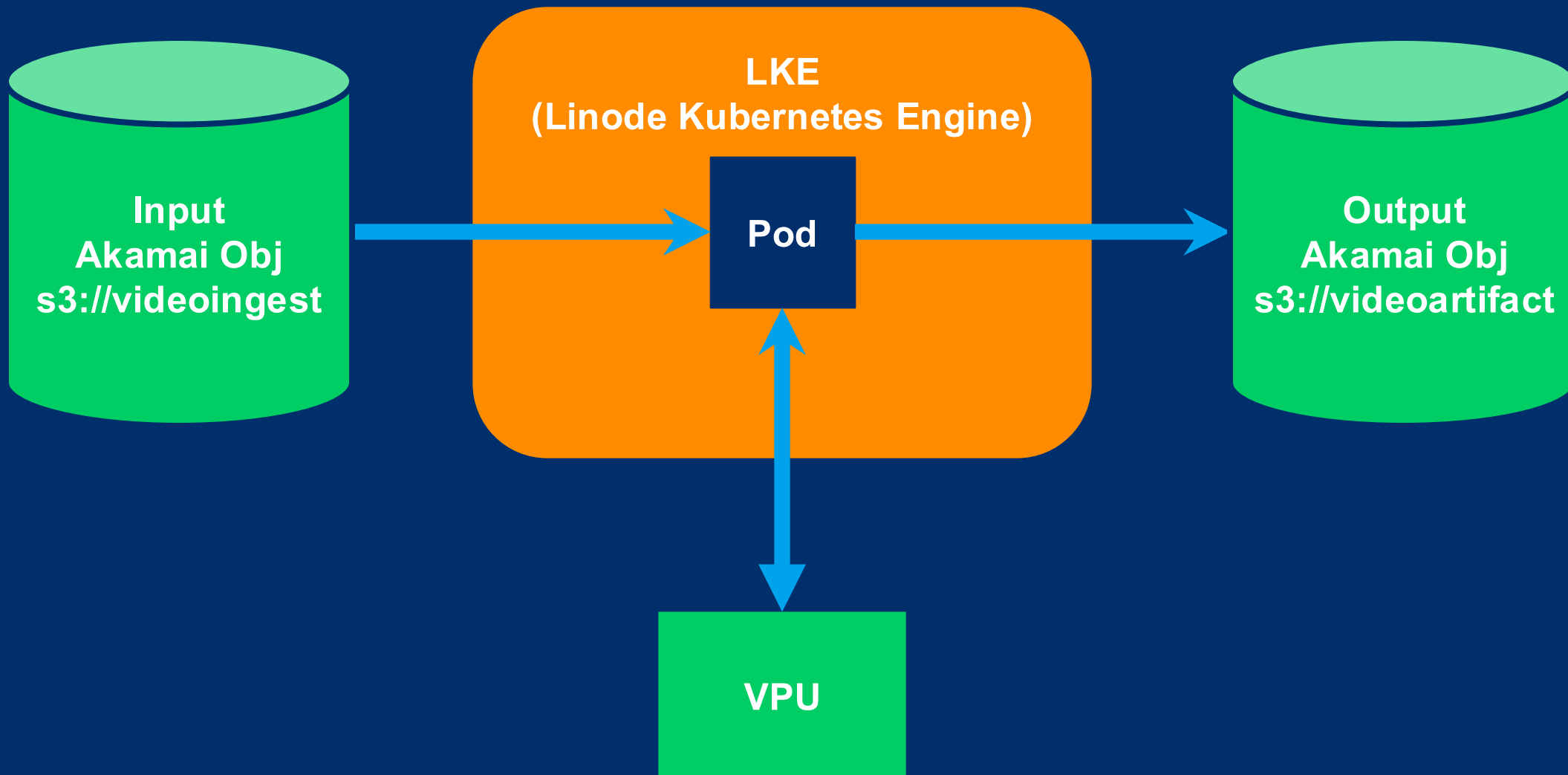
Ease-of-Use

Modern, intuitive, easy to use WebUI

Easy Integration

- On the fly with no required 'ingest' steps.
- Live & TSTV works without any integration
- Match existing URL through Publishing Points

Setting up LKE to do transcoding



Input file:

Video: h264 (High) (avc1 / 0x31637661),
yuvj420p(pc, smpte170m, progressive),
1920x1080, 26236 kb/s

Output: Apple HLS profile AV1+h264+hevc

Resolution	Bitrate (Video Only)
1920×1080	7800 kbps
1280×720	5000 kbps
960×540	2800 kbps
854×480	1400 kbps
640×360	800 kbps
416×234	400 kbps

Cluster Label

MediaCluster

Cluster Tier

Select the cluster tier that's appropriate for your intended workloads. Choose LKE for smaller deployments or LKE Enterprise for a more scalable, enterprise-grade solution.

LKE

Small to medium deployments
Up to 250 nodes, 1000 pods
Shared control plane

LKE Enterprise

Large deployments
Up to 500 nodes, 5000 pods
Dedicated HA control plane

Region

You can use [our speedtest page](#) to find the best region for your current location.



DE, Frankfurt 2 (de-fra-2)



Kubernetes Version

1.33



[Dedicated CPU](#)[Shared CPU](#)[High Memory](#)[GPU](#)[Premium CPU](#)[Accelerated](#)

Some plans do not include bundled network transfer. If the transfer allotment is 0, all outbound network transfer is subject to charges.

[Learn more about transfer costs.](#)

Accelerated instances leverage ASICs to accelerate specialized tasks such as video transcoding, media processing, and other compute heavy workloads. [Learn more](#) about our Accelerated plans.

Plan	Monthly	Hourly	RAM	CPUs	Storage				
NETINT Quadra T1U x1 Small	\$280	\$0.42	16 GB	8	200 GB	—	3	+	Add
NETINT Quadra T1U x1 Medium	\$352	\$0.53	24 GB	12	300 GB	—	3	+	Add
NETINT Quadra T1U x2 Small	\$488	\$0.73	24 GB	12	300 GB	—	3	+	Add

Cluster Summary

NETINT Quadra T1U x1 Small ×
Plan

8 CPUs, 200 GB Storage

—

3

+

\$840.00/month

Additional services added to the cluster
may incur charges.

[See pricing.](#)

\$840.00/mo

Create Cluster

Cluster Label ↓		Version ↓↑		Created ↓↑	Region ↓↑	Total Memory	Total CPUs		
prototype	HA	1.32	UPGRADE	2024-09-19 14:51	SE, Stockholm	96 GB	48 CPUs	Download Kubeconfig	Delete
MediaCluster		1.32	UPGRADE	2025-03-22 19:35	DE, Frankfurt 2	48 GB	24 CPUs	Download Kubeconfig	Delete



Labels and Taints: NETINT Quadra T1U x1 Small Plan ✕

Manage custom labels and taints directly through LKE. Changes are applied to all nodes in this node pool. [Learn more](#).

Labels

Labels are key-value pairs that are used as identifiers. Review the guidelines in the [Kubernetes documentation](#). [↗](#)

Node Label

No labels

Add Label

Label

netint-device: enable

Add

Cancel

Clone NetInt K8S device plugin

```
~$ git clone https://github.com/NETINT-Technologies/netint_k8_device_plugin
Cloning into 'netint_k8_device_plugin'...
remote: Enumerating objects: 76, done.
remote: Counting objects: 100% (76/76), done.
remote: Compressing objects: 100% (53/53), done.
remote: Total 76 (delta 23), reused 60 (delta 13), pack-
reused 0 (from 0)
Receiving objects: 100% (76/76), 46.32 KiB | 5.79 MiB/s,
done.
Resolving deltas: 100% (23/23), done.
```

Install device plugin into cluster

```
~$ cd netint_k8_device_plugin
~$ make deploy-netint
~$ helm install netint deploy/helm/netint

NAME: netint
LAST DEPLOYED: Mon May 19 13:08:01 2025
NAMESPACE: default
STATUS: deployed
REVISION: 1
TEST SUITE: None
```

Add credentials

```
~$ kubectl create secret generic aws-creds \  
  --from-literal=AWS_ACCESS_KEY=[YOUR ACCESS KEY] \  
  --from-literal=AWS_SECRET_ACCESS_KEY=[YOUR SECRET KEY]
```

```
~$ kubectl create configmap s3-buckets \  
  --from-literal=SRCBUCKET=videoingest \  
  --from-literal=DSTBUCKET=videoartifacts
```


EXPLORER

1 unsaved

IBC2025

- vpu
 - docker
 - Dockerfile
 - dot.s3cfg
 - run.sh
 - k8s
 - cronjob.yaml
 - job.yaml
 - pod1.yaml
 - pod2.yaml
 - podwait.yaml
 - Description.txt

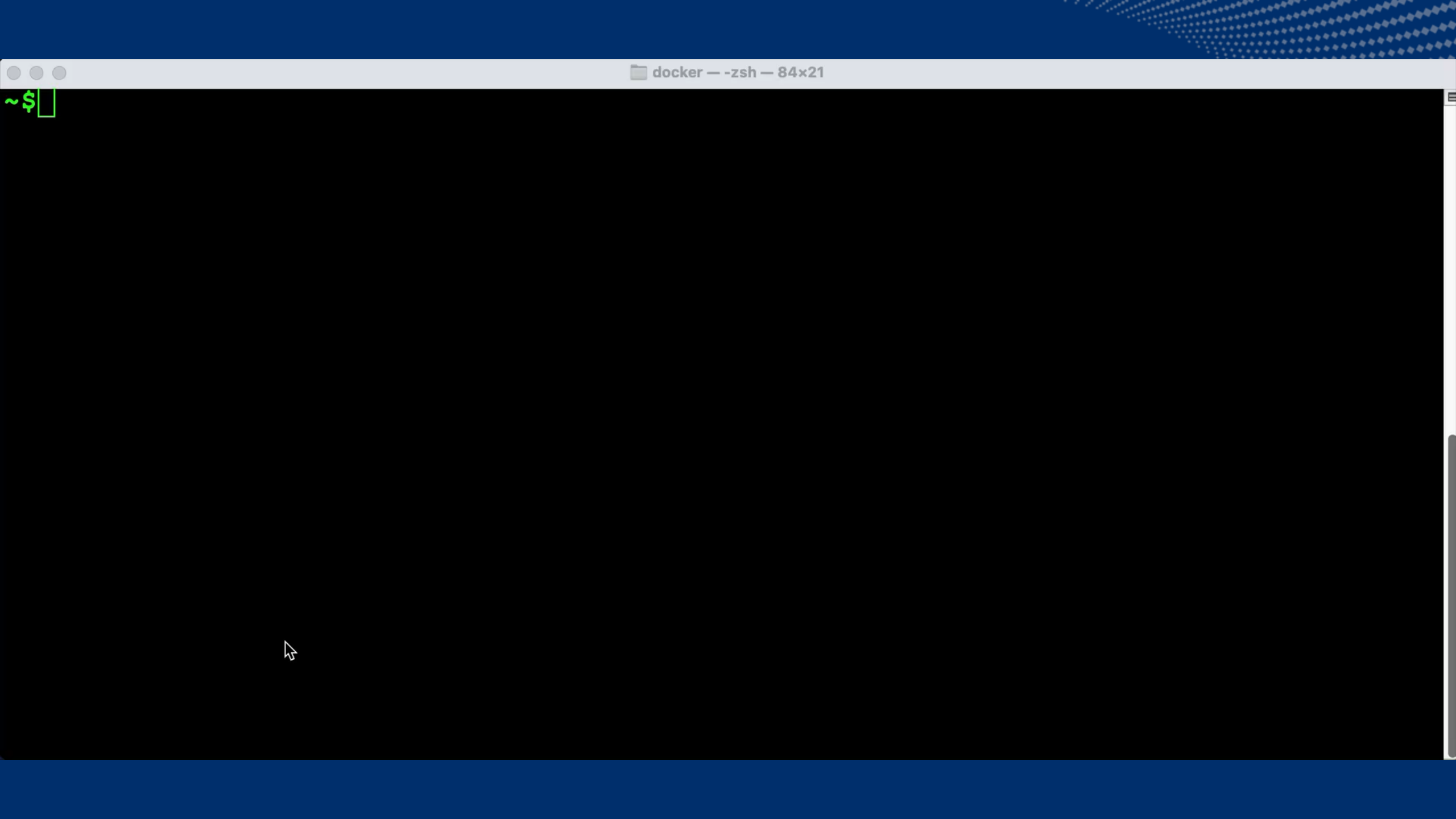
OUTLINE

TIMELINE

yaml pod1.yaml pod2.yaml run.sh podwait.yaml Dockerfile dot.s3cfg

vpu > docker > Dockerfile > ...

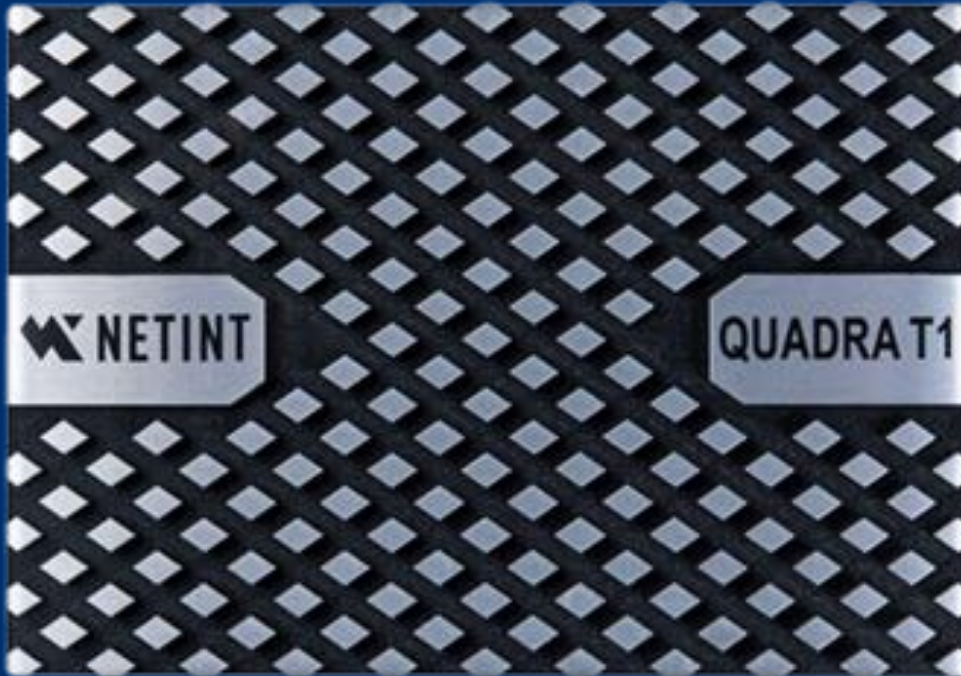
```
1 FROM netint/quadra_ubuntu-24.04_ffmpeg:latest (last pushed 1 month ago)
2
3 RUN apt-get update -y && apt-get -y install s3cmd
4 ADD dot.s3cfg /root/.s3cfg
5 ADD run.sh /root/run.sh
6 RUN chmod 755 /root/run.sh
7
```



~\$



Take a VPU for a spin today!



\$500 Free compute credit!



S SCALSTRM