Service Aware Zero Trust Container Network and Its Offloading to DPU

向阳朝 XIANG Yangzhao
神州数码 DigitalChina
Shanghai, China

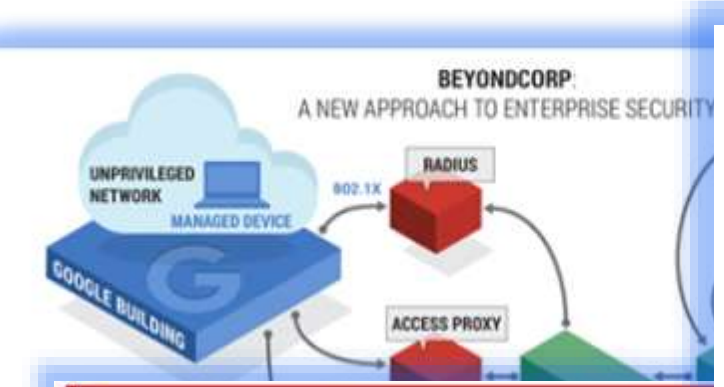
KubeCon + CloudNativeCon + Open Source Summit China 2023

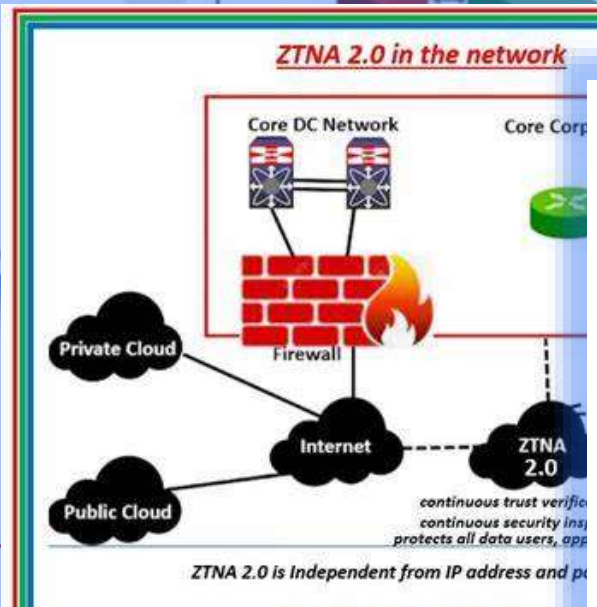
神州数码 Digital China

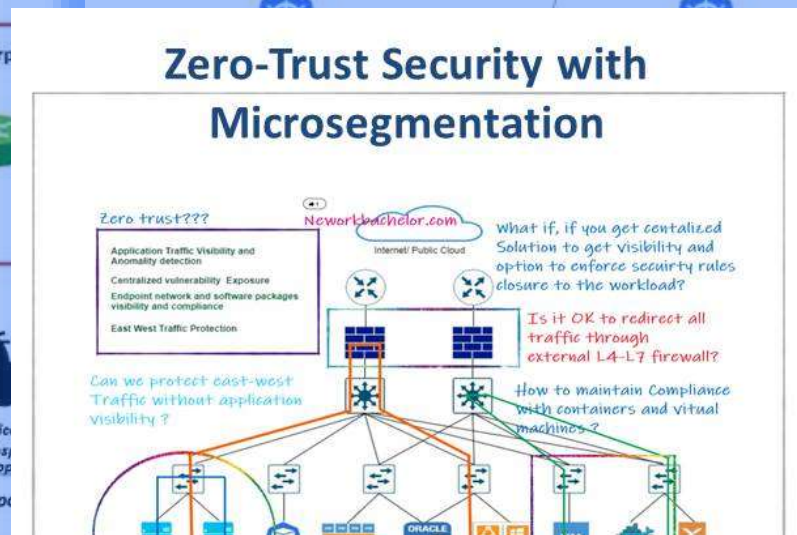
Security Level:

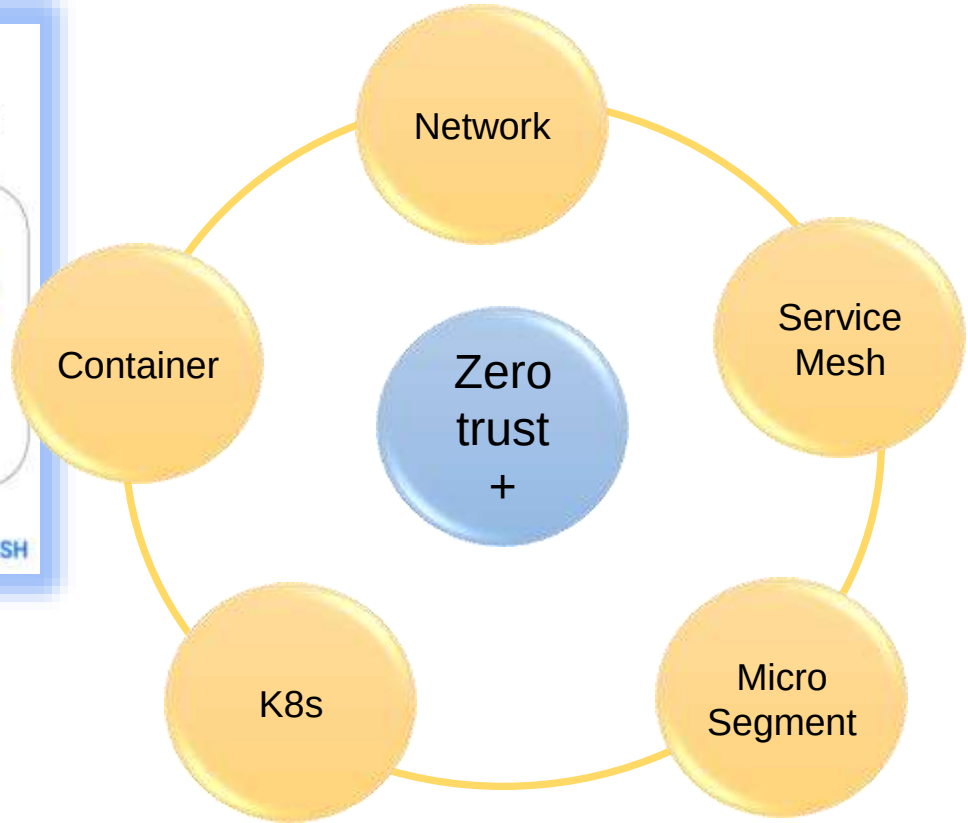
What is zero trust

- does not grant implicit trust to users, devices, and services
- Pwd / certificate is not enough
- Zero trust + everything









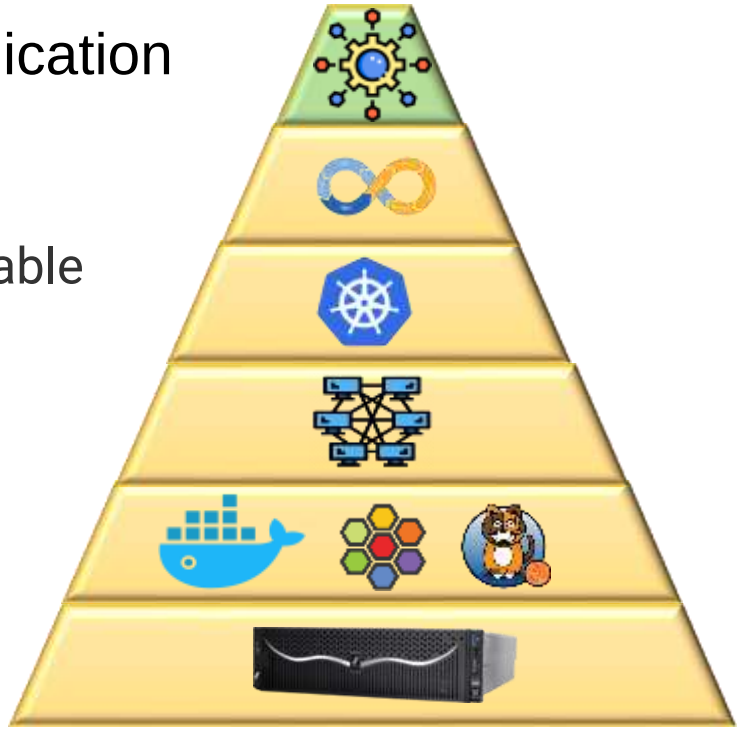
How to achieve zero trust

神州数码 Digital China

- Challenge
  - Identity theft is popular
  - east west lateral movement attack is popular
- Target
  - Limit the effect of identity theft
  - Limit the lateral movement attack
  - Block access based on any entity (user, device, service, location)
- Action--No single component or function is sufficient to implement ZTA
  - continuous verification: behavior, environment, AI/ML
  - all communications are encrypted
  - mutual authentication
  - certificate and key rotation
  - micro-segmentation : network, service
- A complete zero trust security posture may never be fully achieved
  - Performance , usability, price

Cloud Native Zero Trust Architecture

Container network and service mesh provide zero trust communication



- Cloud Native System
  - loosely coupled systems that are resilient, manageable, and observable
- Cloud Native Application
  - a collection of small, independent, and loosely coupled services
- Cloud Native Infrastructure
  - containers, network, orchestration , CI/CD
  - service mesh

- Action
  - Continuous verification
  - Encrypted communications
  - Mutual authentication
  - Certificate and key rotation
  - Micro-segmentation

→

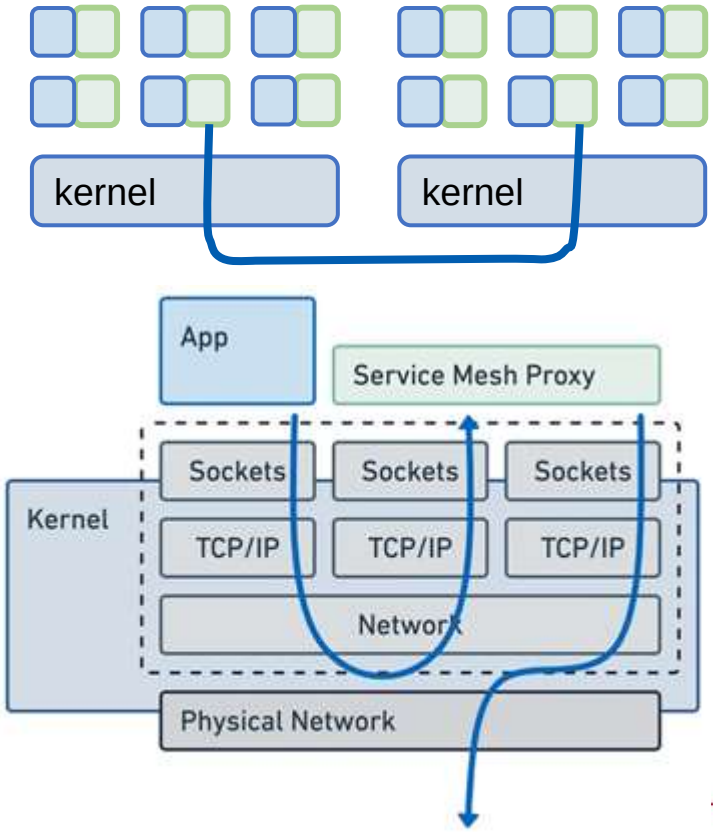
- Telemetry: container, network, service
- Session encryption
- Mutual service authentication
- Certificate and key rotation
- Microservice segmentation

Sidecar & sidecarless service mesh

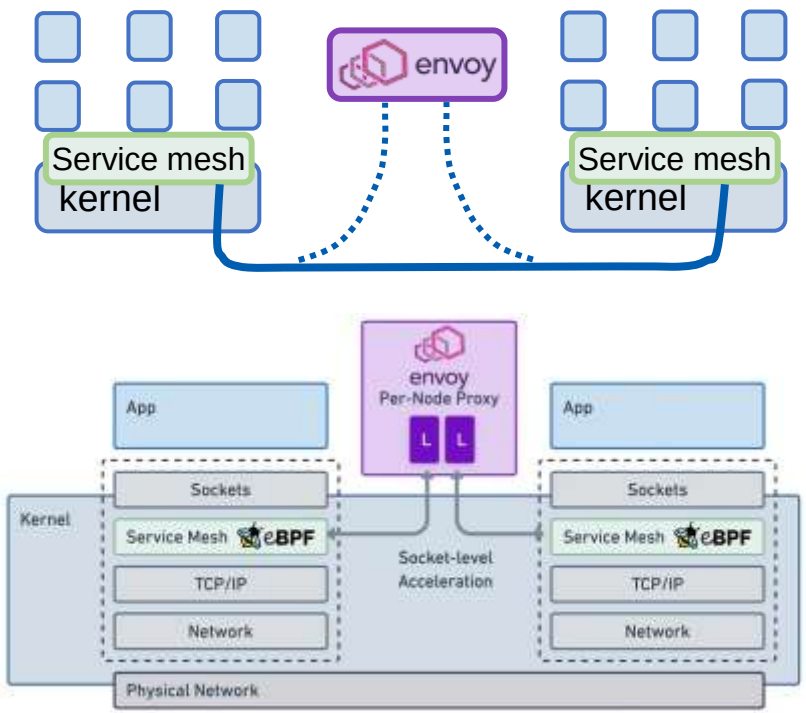
Move L4 to CNI
Decrease proxies
eBPF
Per node/service share

Reduce invasion
decouple POD and mesh infra
reduce latency
Reduce host usage

Istio

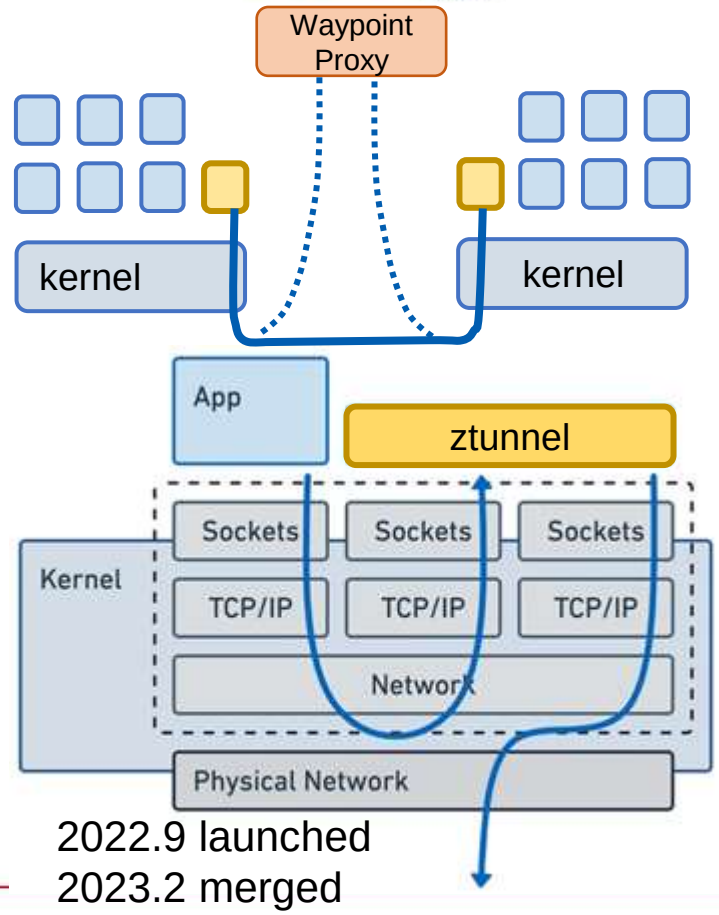


cilium



2022.7 v1.12

ambient mesh



2022.9 launched
2023.2 merged

No perfect solution

神州数码 Digital China

Balance among performance, scalability, security

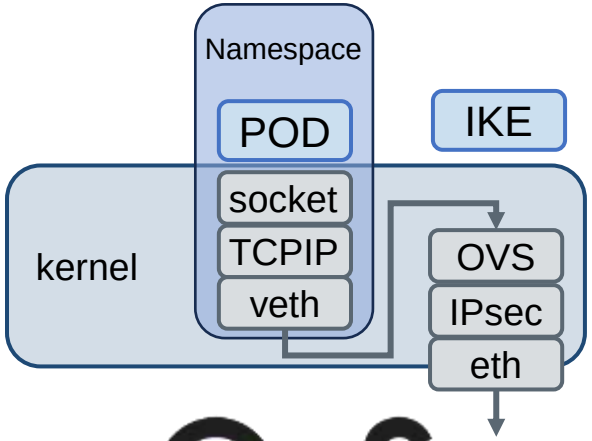
	Istio	cilium	ambient mesh
Latency	★★	★★★	★★
Host CPU usage	★★	★★★★	★★★★★
Host memory usage	★★	★★★★	★★★★★
POD invasion	★★	★★★★★	★★★★★
Kernel Req.	★★★★★	★	★★★★★
Istio compatibility	★★★★★	★★	★★★★
scalability	★★★★★	★★	★★★★
Service awareness	★★★★★	★★★	★★★
Breach radius	★★★★★ POD	★★ Node	★★ Node
vulnerability	★★ App, Envoy	★★★★ Envoy	★★★★ Ambient

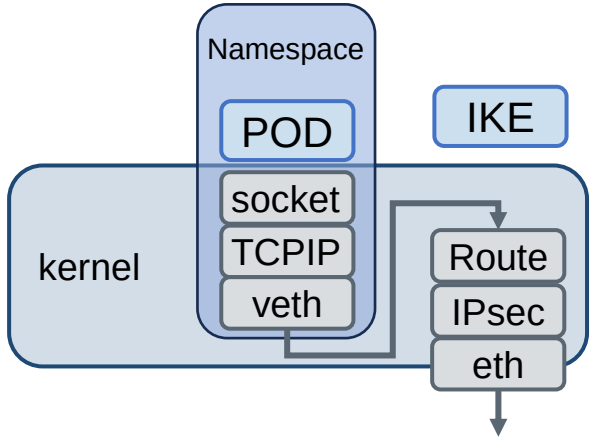
More stars is better

Zero trust container network

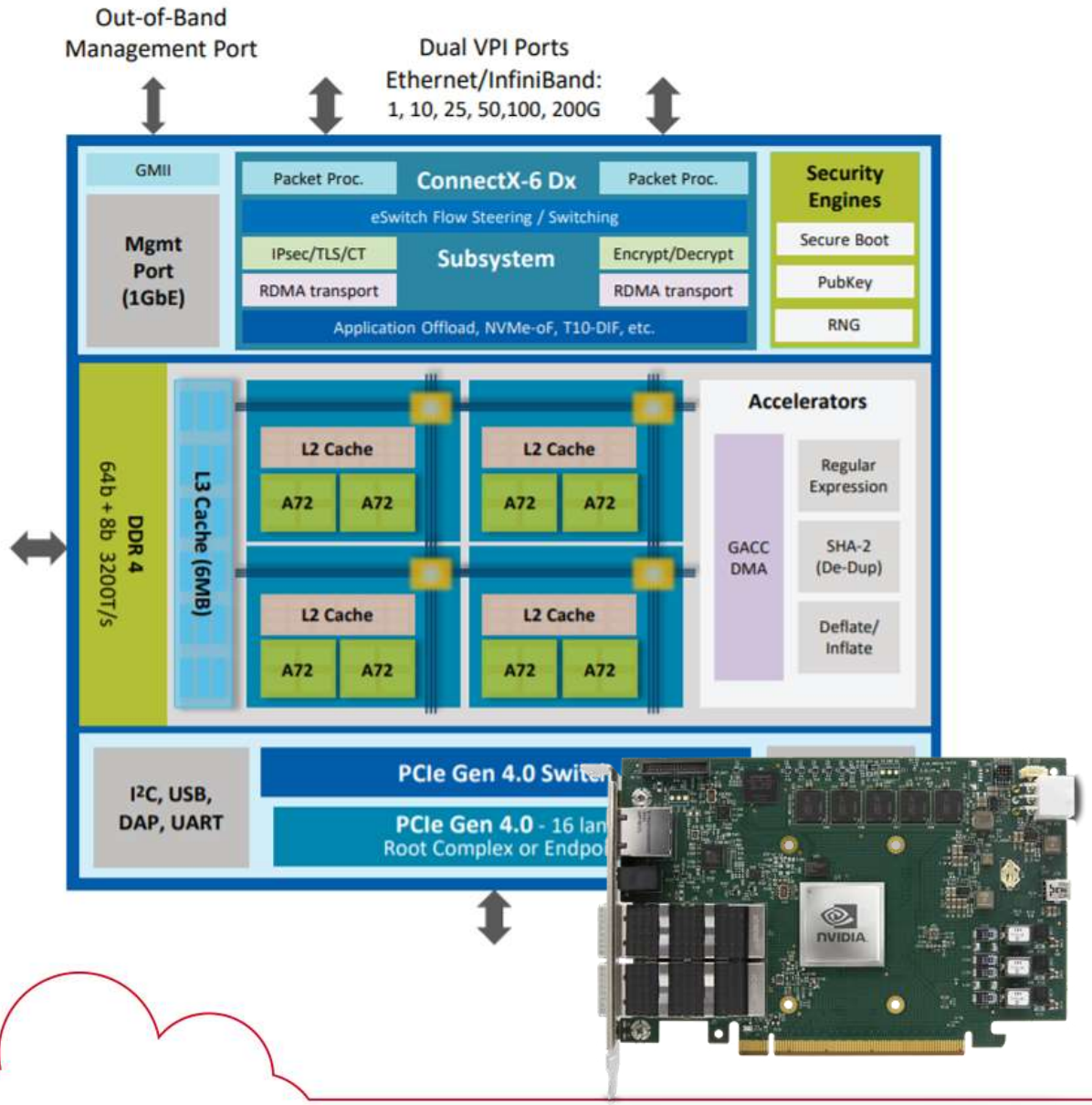
Service mesh is not deployed in many scenarios, service awareness is needed for service identity authentication in container network tunnel.

- packets are transported from one node to another via IPsec/wireguard tunnel.
- do not currently provide any support for IPsec encryption for traffic not encapsulated in a tunnel
- Workload identity is based on IP.





What does DPU provide



- Overlay network acceleration
  - VXLAN, GENEVE, NVGRE
- Connection tracking
- Flow sampling and statistics
- IPsec/TLS data-in-motion encryption
  - AES-GCM 128/256-bit key
- Public key accelerator (PKA)
  - RSA, Diffie-Hellman, DSA, ECC,EC-DSA, EC-DH

Zero trust cloud native with DPU

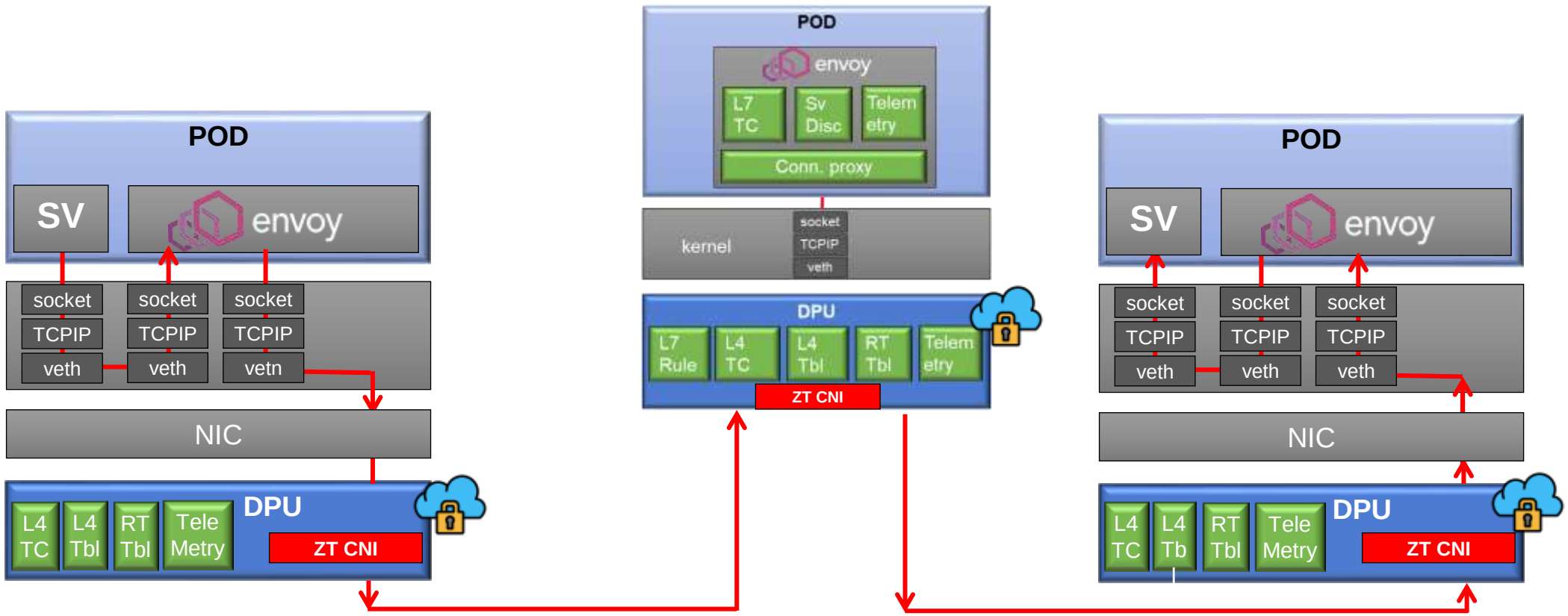
- Targets
  - Increase throughput, decrease latency
  - Compatible with istio eco.
  - Keep or improve security
  - Support scenarios w/o sidecar
  - Decouple app and infra

- Challenges of offloaded ZT CNI
  - Service aware
  - Breach radius

Zero trust cloud native with DPU

神州数码 Digital China

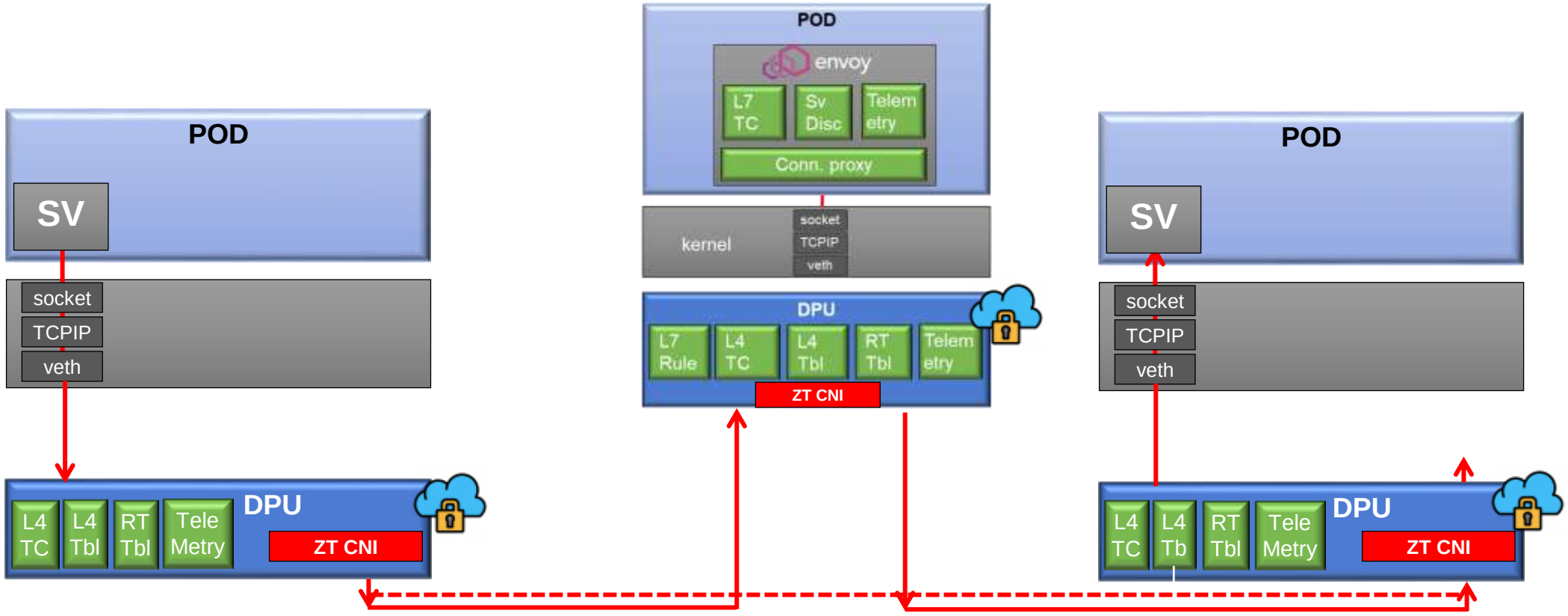
- Targets
  - Increase throughput, decrease latency
  - Compatible with istio eco.
  - Keep or improve security
  - Support scenarios w/o sidecar
  - Decouple app and infra
- Challenges of offloaded ZT CNI
  - Service aware
  - Breach radius



Zero trust cloud native with DPU

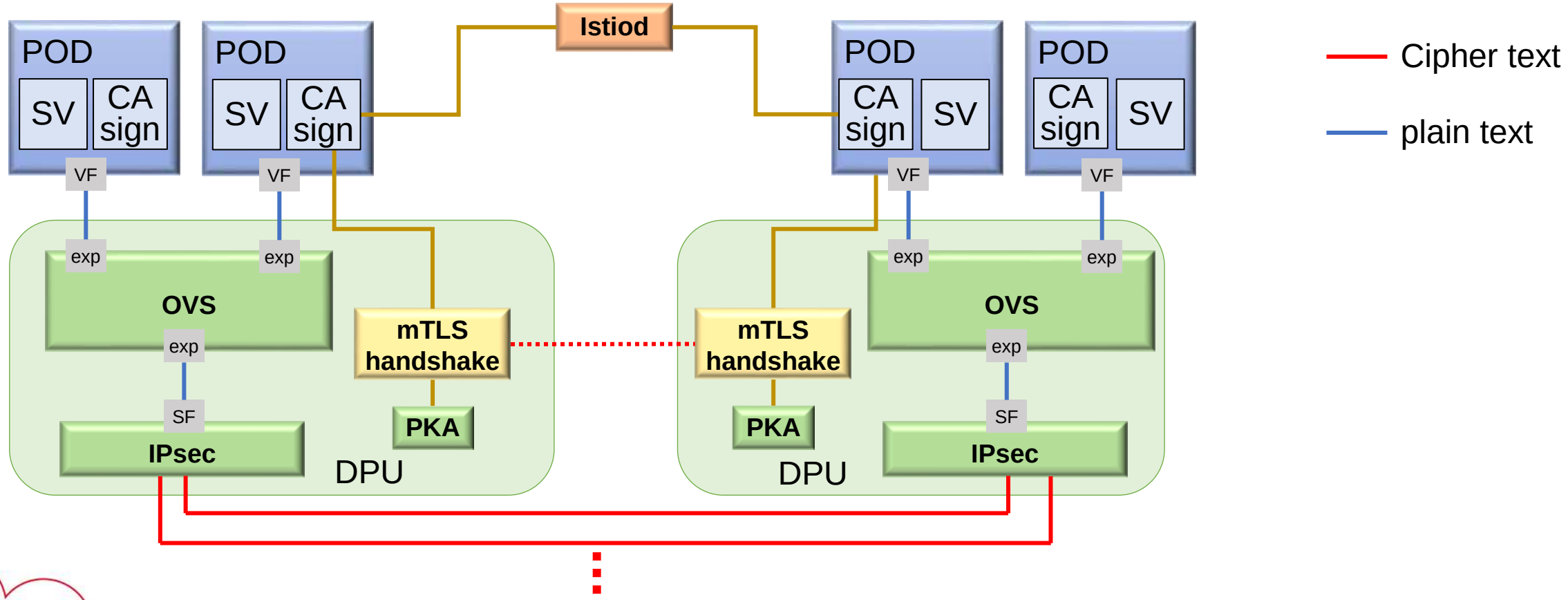
- Targets
  - Increase throughput, decrease latency
  - Compatible with istio eco.
  - Keep or improve security
  - Support scenarios w/o sidecar
  - Decouple app and infra

- Challenges of offloaded ZT CNI
  - Service aware
  - Breach radius



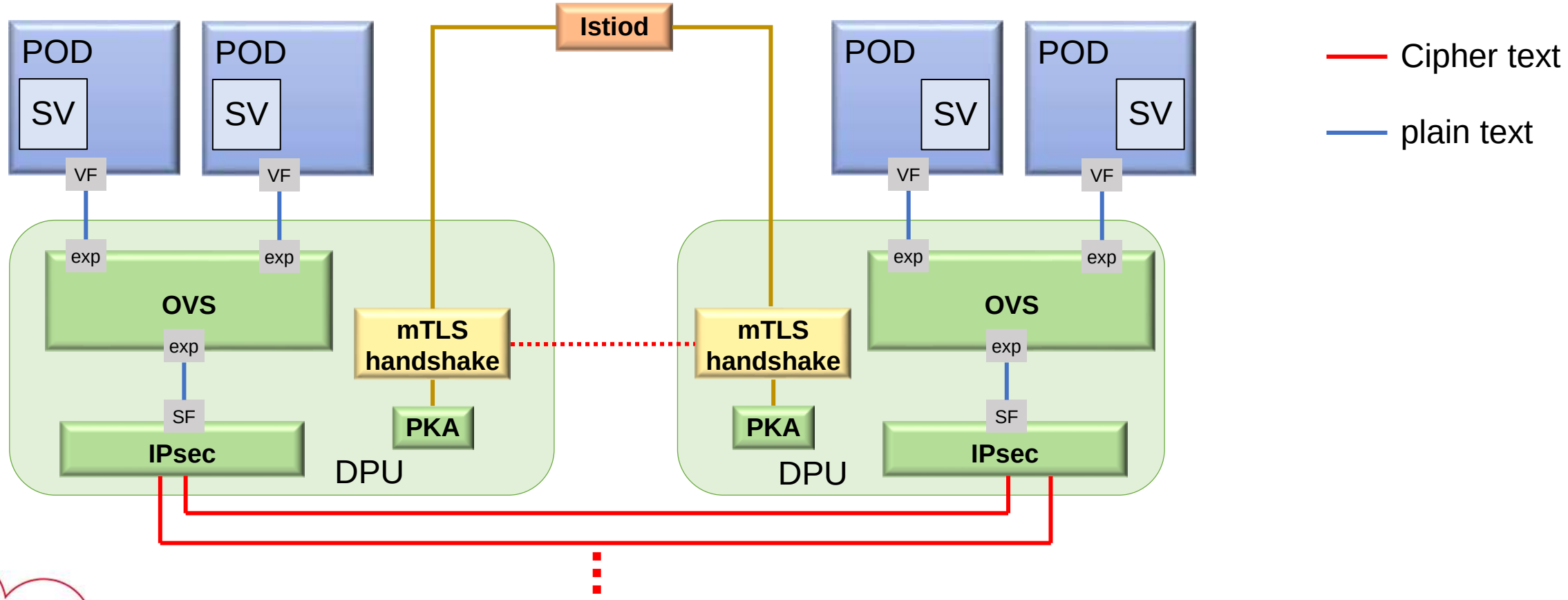
Architecture of zero trust CNI on DPU

- Per POD IPsec tunnel
- mTLS handshake for each tunnel
- Use vfid in metadata for POD identification
- Boost signature performance with larger breach radius



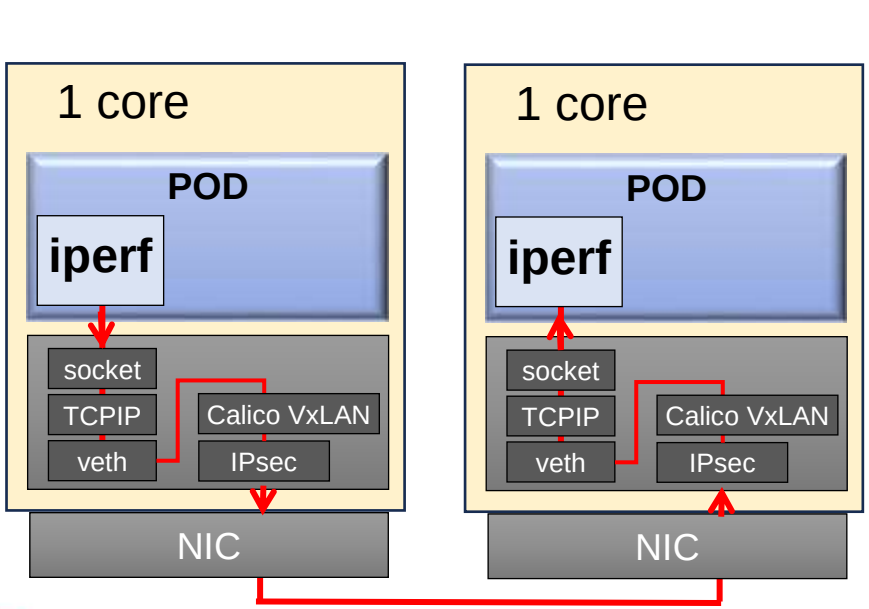
Architecture of zero trust CNI on DPU

- Per POD IPsec tunnel
- mTLS handshake for each tunnel
- Use vfid in metadata for POD identification
- Boost signature performance with larger breach radius



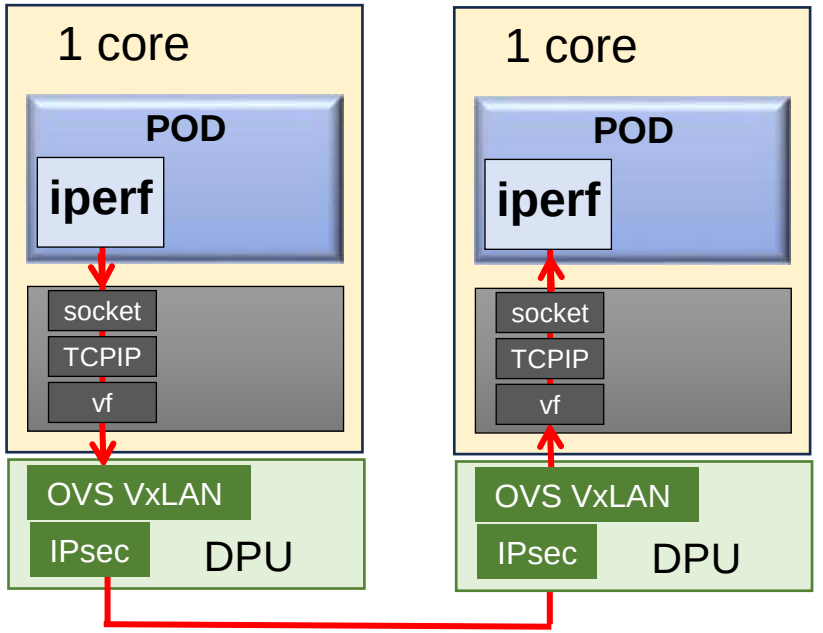
Testbed of ZT CNI w/wo DPU

- Testbed
  - server: DigitalChina KunTai 2280, CPU: KunPeng 920, 2*48 cores @2.6G
  - DPU: nVidia bluefield 2
  - TSO and Armv8 Cryptographic Extension are enabled by default
  - Iperf sending 4~8 tcp streams is bond to 1 core
- 1.8x~23.5x encrypted TCP throughput improvement @ 1 core



Calico+VxLAN+ipsec

vs

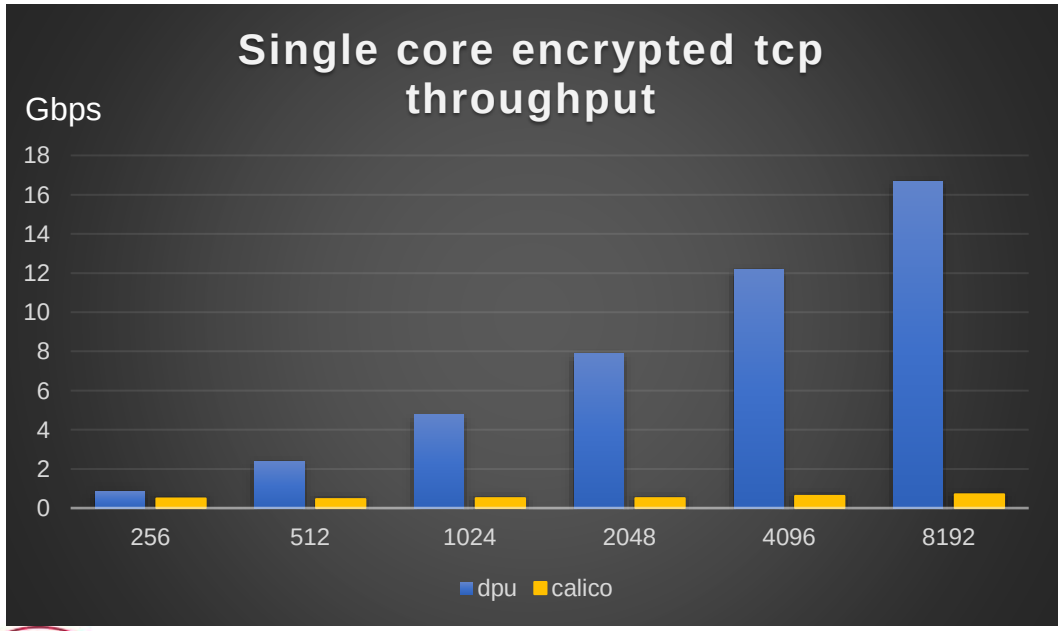


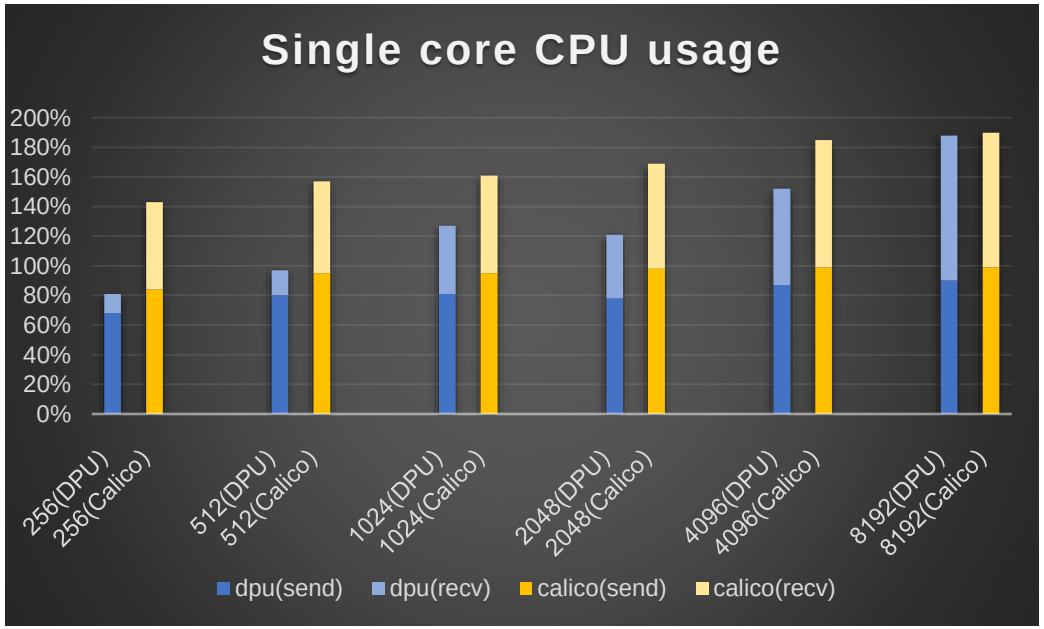
DPU ovs+VxLAN+ipsec

Performance of ZT CNI w/wo DPU

神州数码 Digital China

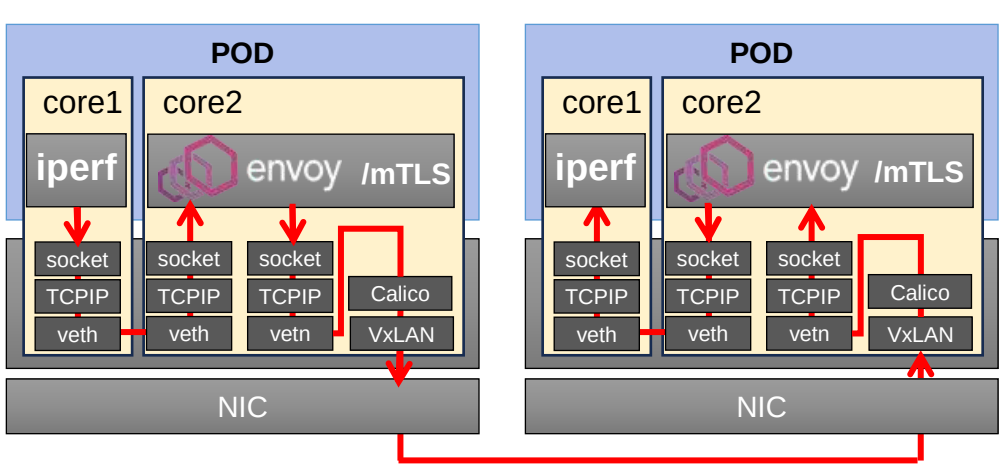
- 1.8x~23.5x encrypted tcp throughput improvement
- 40~1% CPU utilization reduce
  - Iperf uses 10~30% CPU
  - 3x~24x throughput improvement per core





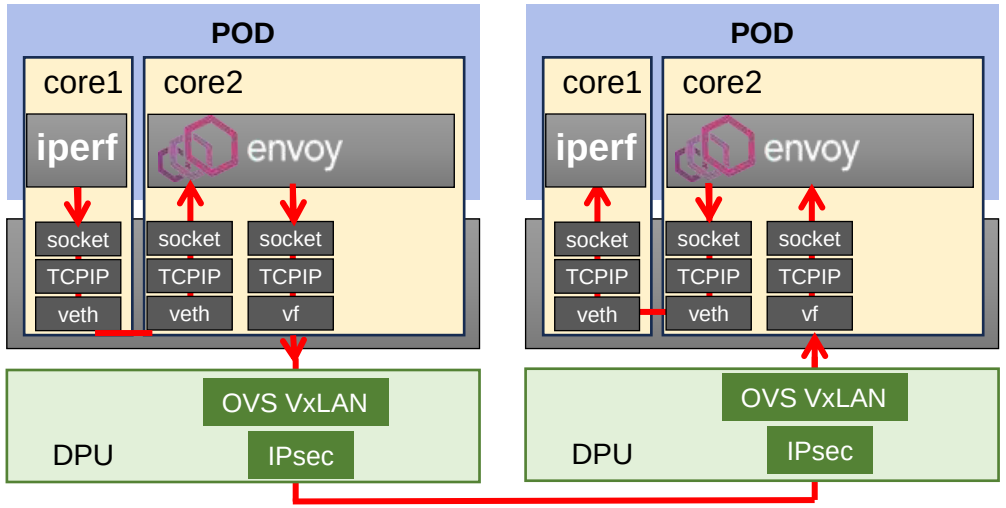
Testbed of sidecar w/wo ZT DPU CNI

- Testbed
  - server: DigitalChina KunTai 2280, CPU: KunPeng 920, 2*48 cores @2.6G
  - DPU: nVidia bluefield 2
  - TSO and Armv8 Cryptographic Extension are enabled by default
  - Iperf sending 4 tcp streams is bond to 1 core
- 3.2x~4.9x encrypted TCP throughput improvement @ 1 core
- 1.5x~4.1x encrypted TCP throughput improvement @ 4 cores



Envoy sidecar+mTLS+Calico+VxLAN

VS



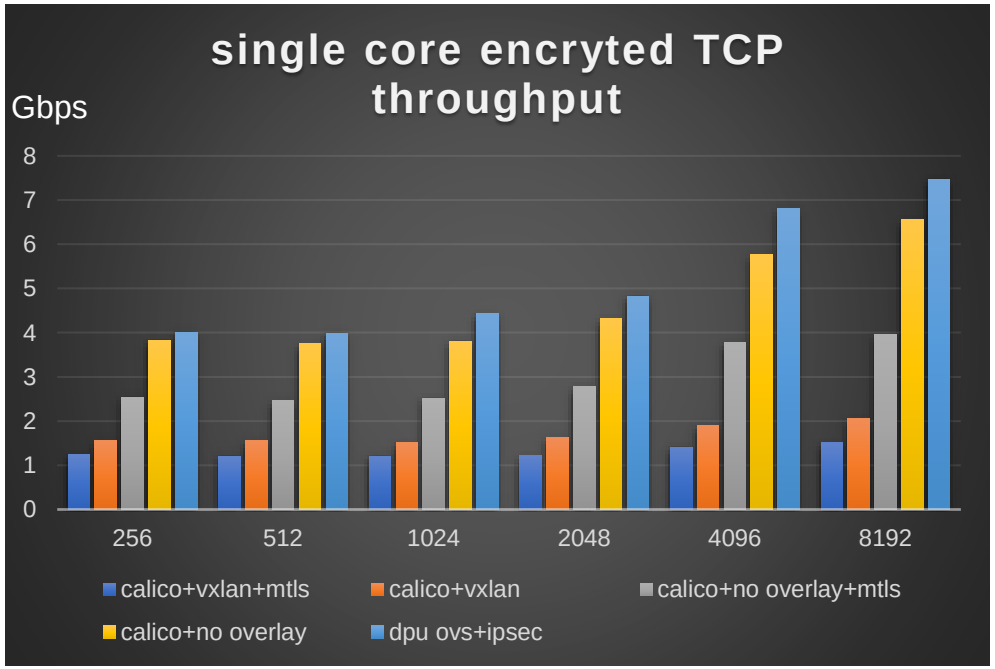
Envoy sidecar+DPU (ovs+VxLAN+ipsec)

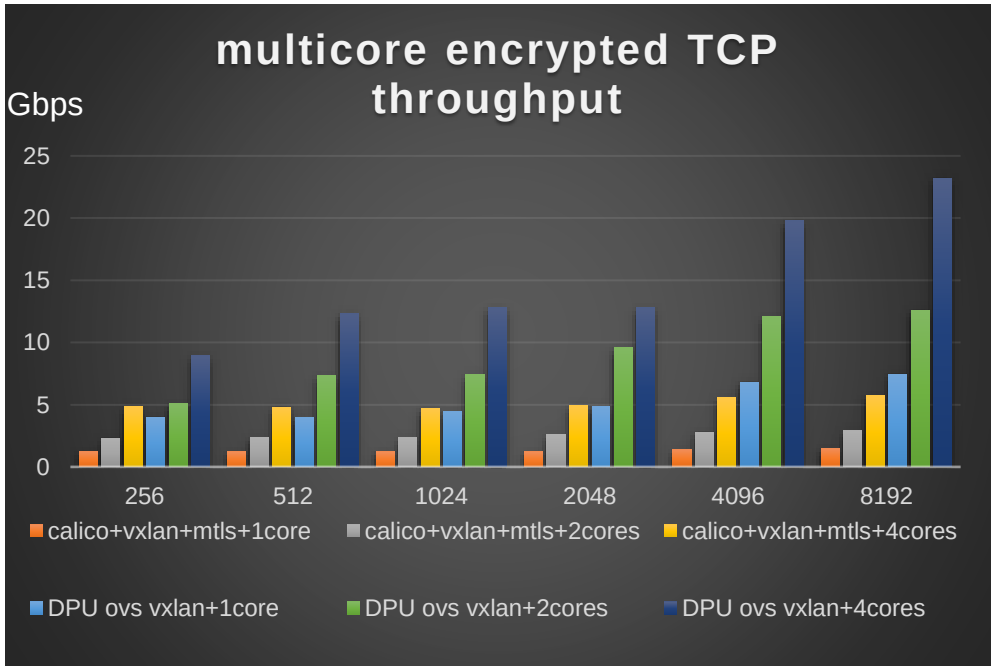
Throughput of sidecar w/wo ZT DPU CNI

- vxlan reduces throughput by 60~70%
- mtls reduces throughput by 30~40%

- 3.2x~4.9x improvement @ 1 core
- 1.5x~4.1x improvement @ 4 cores

TSO(vxlan) has greater impact than mtls
Kernel network stack locks have impact on multicore

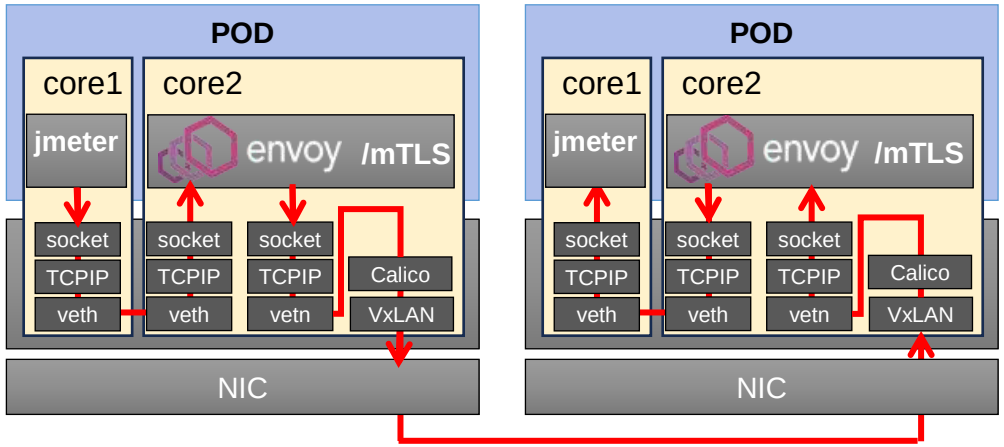




Higher is better

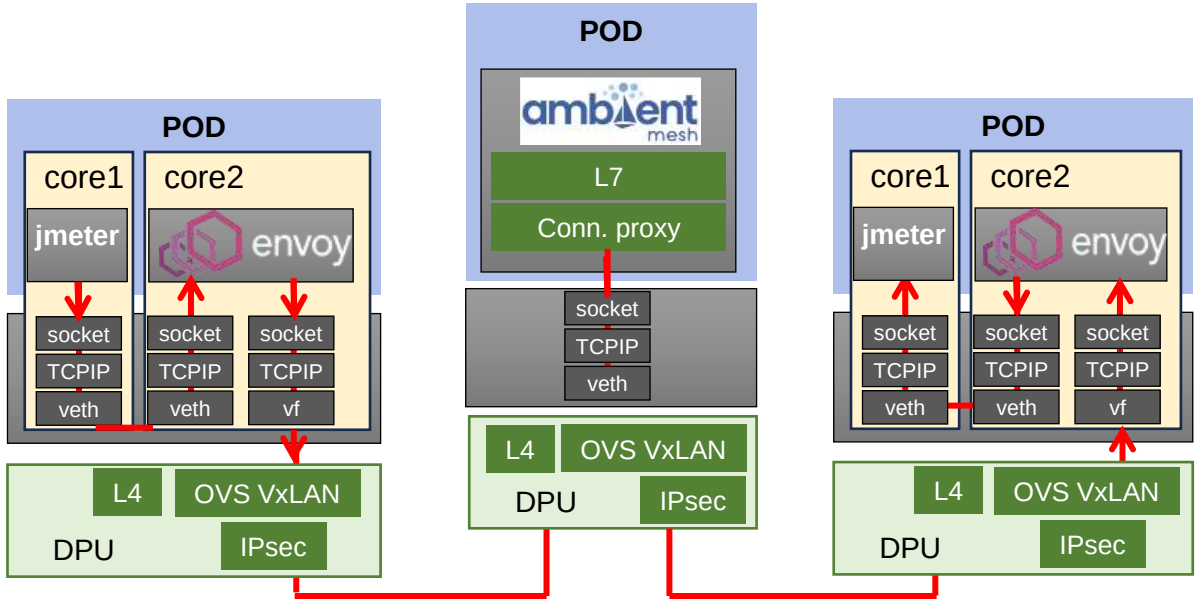
Testbed of service mesh w/wo sidecar

- Testbed
  - server: DigitalChina KunTai 2280, CPU: KunPeng 920, 2*48 cores @2.6G
  - DPU: nVidia bluefield 2
  - TSO and Armv8 Cryptographic Extension are enabled by default
  - Jmeter sends 50 concurrent http streams
- 4x encrypted http QPS improvement per 1 core
- 50% encrypted http latency reduce



envoy+mTLS+Calico+VxLAN

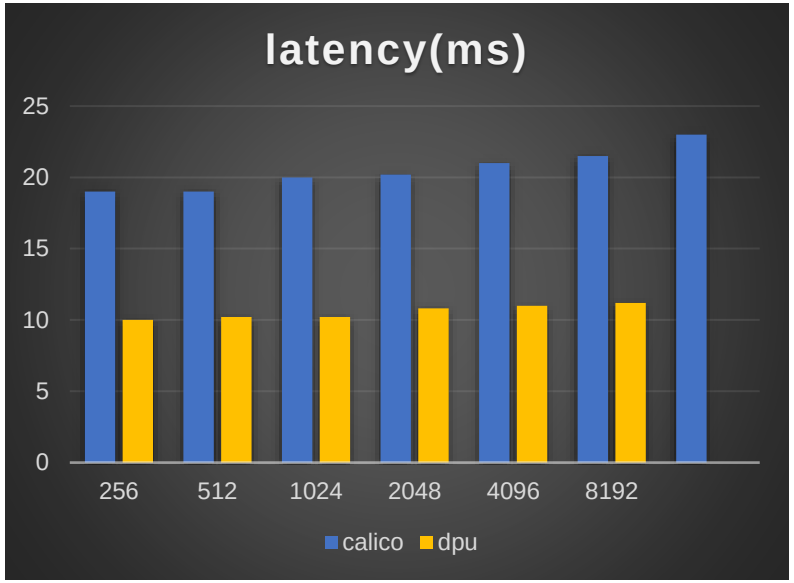
VS

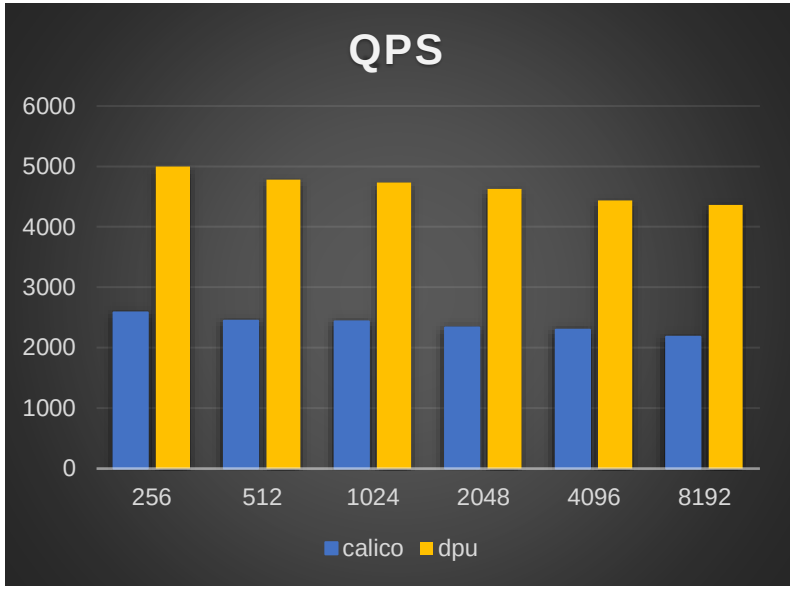


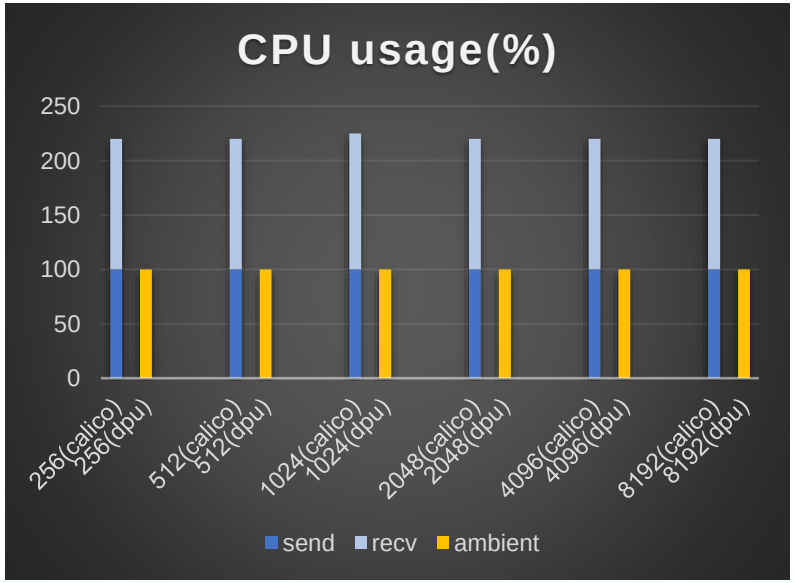
ambient+DPU ovs+VxLAN+ipsec

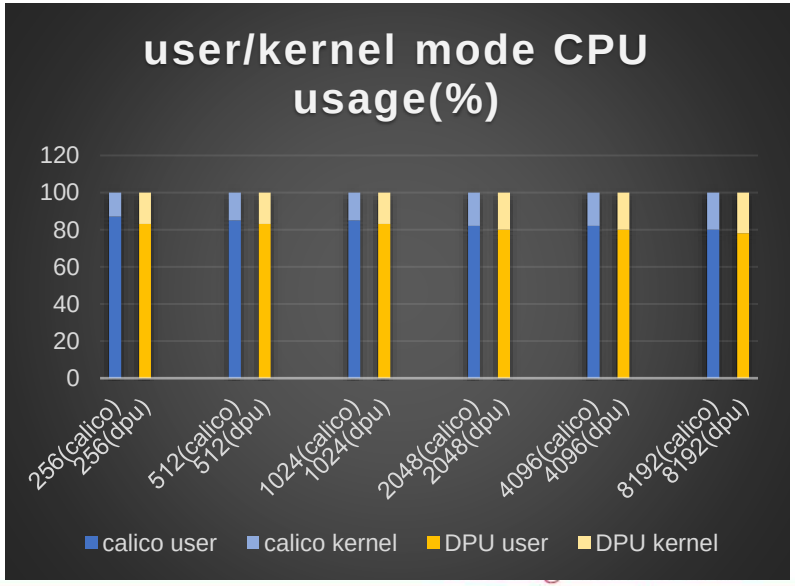
Performance of service mesh w/wo sidecar

- 2x encrypted http QPS improvement
- 54% CPU usage reduce
  - 4x QPS improvement per core
- 50% latency reduce
- 80% CPU is used by envoy L7
  - L7 parse
  - Rule match
  - Further improvement direction









KUNTAI 神州鲲泰

智 算 神 州
鲲 泰 领 航

神州数码
Digital China