Predictable Performance through Prometheus and Topology Aware Scheduling

Killian Muldoon, Intel
Tom Golway, Hewlett Packard Enterprise

Rise of the Digitally Aware Enterprise

- Digital Transformation is not about automating business processes
- Cultural shift toward business innovation that is digitally aware
- Business ecosystem is becoming dynamic and real-time
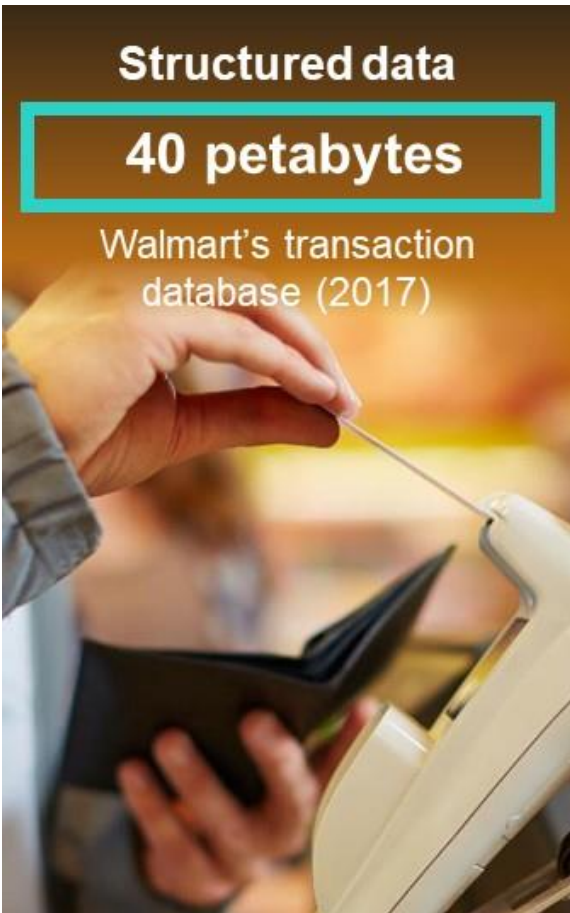- Expectations are rapidly changing, plasticity is the new agility

Shifting Consumption Models

System of record	System of engagement	System of action

Structured data

40 petabytes

Walmart's transaction database (2017)

Human interaction data

4 petabytes

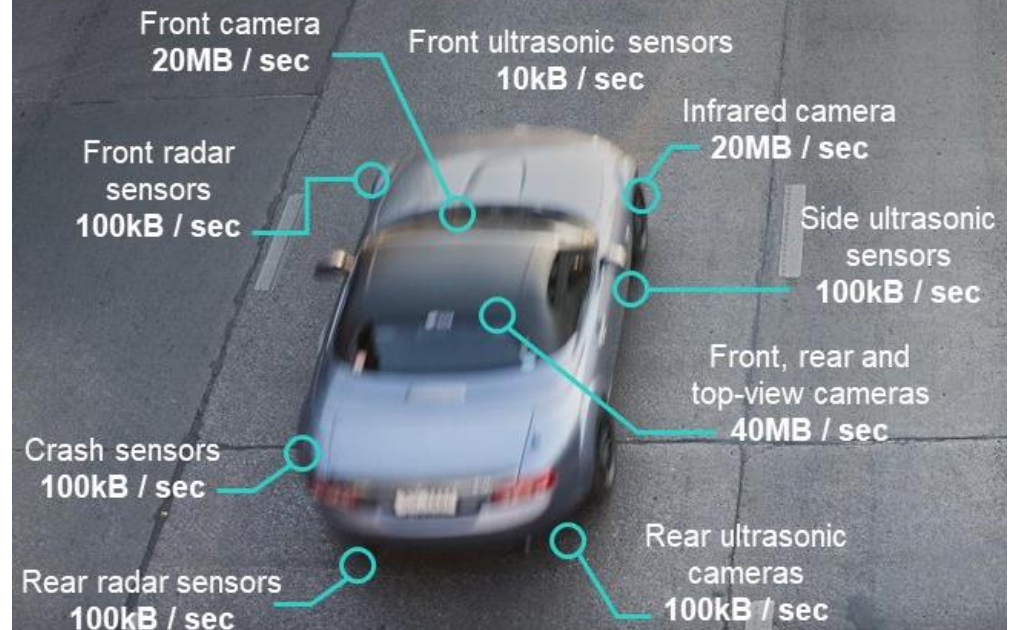Per-day posting to Facebook across 1.1 billion active users (May 2016)

4kB per active user

Digitization of Analog Reality

4 petabytes a day

10m self-driving cars by 2020

Front camera 20MB / sec

Front ultrasonic sensors 10kB / sec

Infrared camera 20MB / sec

Front radar sensors 100kB / sec

Side ultrasonic sensors 100kB / sec

Front, rear and top-view cameras 40MB / sec

Crash sensors 100kB / sec

Rear radar sensors 100kB / sec

Rear ultrasonic cameras 100kB / sec

- Volume of data is projected to grow exponentially.

- Variety of data is increasing.

- Sources of data creation and/or ingestion to the application ecosystem has grown proportionally to the increased growth of IoT.

- The value and usage of the data is changing.

- Data has gravity which influences where compute needs to be located.

- Cloud Native has created a significant challenge for the veracity of data.

- App architectures moving toward disaggregation models with a governing process managing the flow of execution

- App architectures will use a composable model with a mix of core code, libraries, external apps, and microservices

- Individual components of the App will be executed on the platform that best fits the business process Class of Service

- App topologies will range from a simple tree structure to a more complex distributed mesh

- Overall App service levels dependent on individual µservice metrics such as latency and jitter.

- Apps require underlying infrastructure to support the concept of plasticity

Hardware Traits

- Kubernetes Infrastructure knows what it's made of, but not what it can do
- Predictable performance means deep knowledge of specific hardware
- App developers shouldn't need to learn the layout of every system in a cluster!
- Need to move from components to capabilities:

Components model:
Request resources directly in pod spec
Device Plugins and Node Feature Discovery

Capabilities model:
Request performance/other characteristics
Traits engine matches application and node traits

Pod Requirements:
9 CPU cores @ 3.5 GHz +
8 Net Virt Functions
@10GB/s
10 Memory Huge Pages
Power headroom of 150W
Numa alignment

Pod Requirements:
10GB Network@1Mpps

Prometheus & Telemetry Scheduling

- Prototype node trait discovery & matching engine using Prometheus and Telemetry Aware Scheduling

- Prometheus converts raw telemetry (Collectd, Node Exporter, NFD etc.) to platform capabilities

- Telemetry Aware Scheduling used as a matching engine – filtering out nodes that can't meet the performance guarantees

- Pushes complexity for managing hardware to a new component – matches application class of service model

- Memory Topology Scheduling managed by Prometheus – imperfect solution
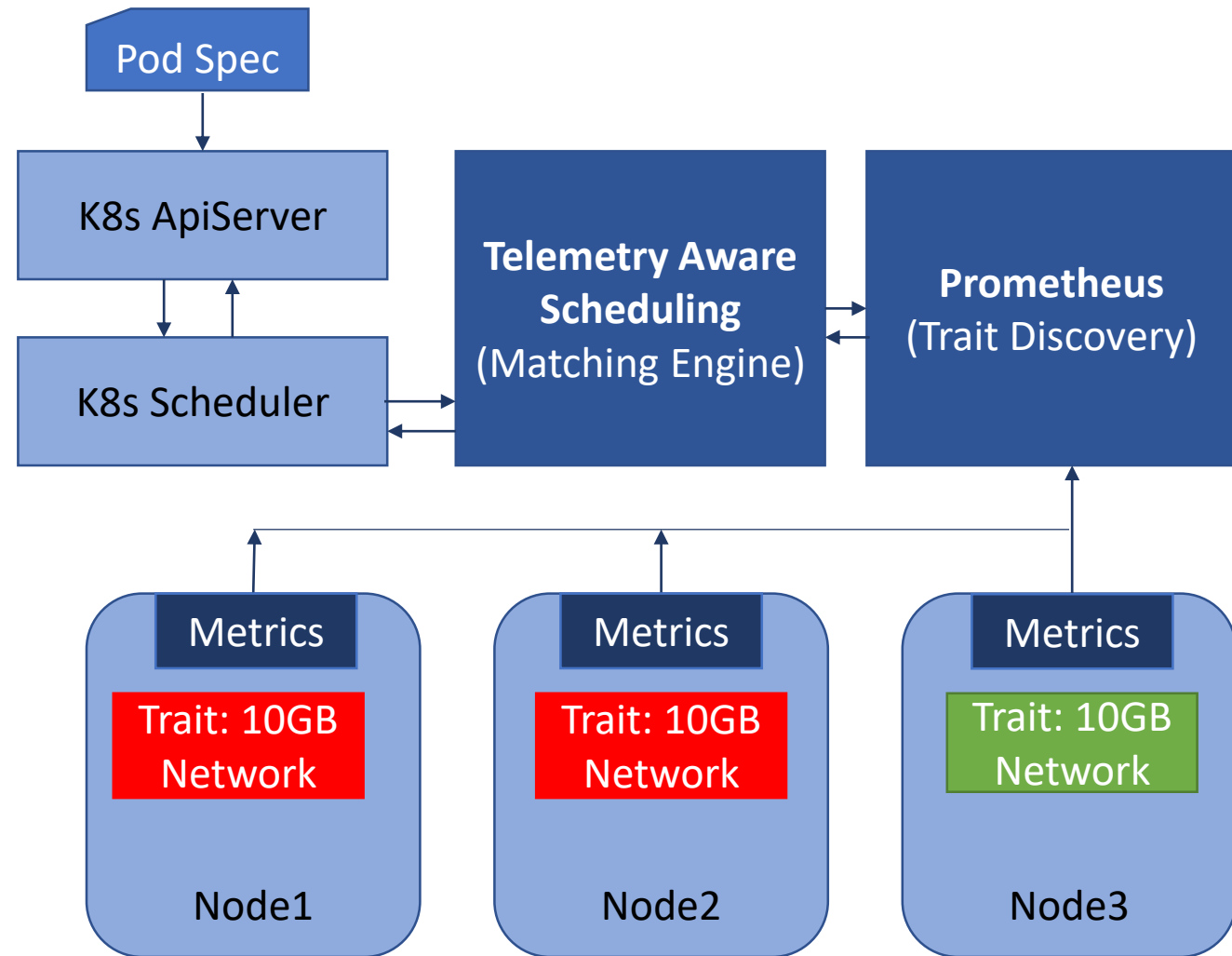
This page intentionally left blank

A new path but Kubernetes isn't ready

- New way to think about class of service for workloads will need a matching effort on the side of hardware

- Increasing hardware complexity with xPU will make it increasingly difficult to correctly configure workload resource configurations

But Kubernetes isn't ready!

- Numa awareness in control plane is #1 roadblock for predictable performance

- More complex platforms mean more uneven topology performance in future

- Work is being done – shout out to Sig Node Resource Management which is working to solve some of the big gaps for HPC & Networking in Kubernetes.