RAN Intelligence Enabling

OCP & Intel® Smart Edge Open

Intel® Network Builders Webinar Aug 23rd 2022

Michael Recchia, Global Telco Solutions Architect, Red Hat
Federico Rossi, Senior Solutions Architect, Red Hat
Hassnaa Moustafa, Principal Engineer, Intel
Sunku Ranganath, Customer Solutions Architect, Intel

Notices and Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

No product or component can be absolutely secure.

Your costs and results may vary.

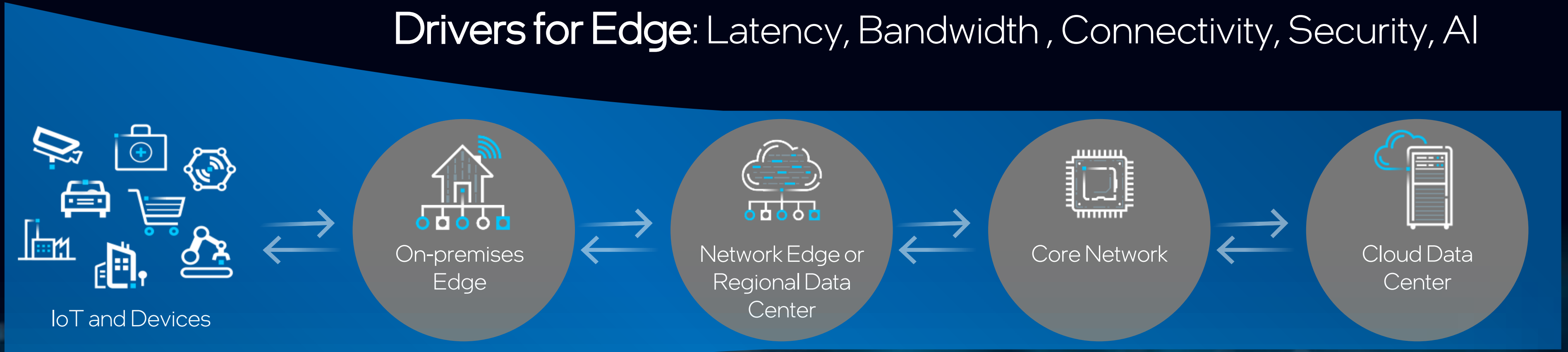
Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Product plans, dates, and specifications are preliminary and subject to change without notice.

Delivering Cloud Native Platforms for the Edge

Drivers for Edge: Latency, Bandwidth , Connectivity, Security, AI



	IoT and Devices	On-premises Edge	Network Edge or Regional Data Center	Core Network	Cloud Data Center
Latency Expectation	Varies <1 ms	<5 ms	<10-40 ms	<60 ms	~100 ms
Bandwidth Expectation		1000+ Mb/s		100 Mb/s	10 Mb/s

Opportunity @ the Edge by 2025

- Multi-access Edge and Private Wireless Hardware, Software, and Services - $29B [1]
- 75% of Data Created Outside Central Data Centers [2]

Key Technology Inflections

- Cloud Native Software
- Connectivity (5G, Multi-Access)
- Artificial Intelligence

Edge of the Future

- Real Time/Deterministic
- On-Demand/Dynamic
- Energy Efficient/Sustainable
- Massively Geo-Distributed at Scale

Lower TCO with a consistent cloud native platform approach across edge locations

[1] MEC definition here refers to MEC2.0 hyperconverged edge. Source: IDC, Omdia, Intel Judgment.

[2] What Edge Computing Means for Infrastructure and Operations Leaders, Gartner, Oct 3, 2018.

What is Intel® Smart Edge?

certified kubernetes

Edge-native Kubernetes Certified Distributed Computing Platform

- Enables deployment and management of container-based workloads with cloud-like ease, resiliency and security at the edge.
- Runs demanding workloads like AI, media, and software-defined networking functions, powered by pre-validated blueprints and solutions provided by Intel and a robust partner ecosystem.
- Powers diverse use cases across industries and delivers performance, security, manageability and sustainability.

Intel Smart Edge for Developers
Develop with us

- Developer Experience Kits* for Edge Video Analytics, ORAN xApps
- Apache 2.0 license
- Community Support
- Available via Github, Intel® Developer Catalog & Intel® Dev Cloud

Intel Smart Edge for Builders
Build with us

- Builder Experience Kits* for Private Wireless, SASE
- Intel royalty-free license
- Intel Basic Support
- Available via Github

Intel Smart Edge for Enterprises
Buy from us

- Turn-key commercial software for private wireless: Edge Controller and Edge Node
- Intel paid license
- Intel Premium Support
- Available via Github

Partner Market Ready Solutions, Partner Commercial Applications, Intel Reference Applications

Consumable as integrated platforms or composable building blocks

Intel® Smart Edge Microservices-based Architecture

Modular Approach to Build Edge Solutions

Developer Outreach

Intel Developer Catalog

Intel Dev Cloud (soon)

Intel® Developer Catalog

One stop-shop for SW access

Edge Reference Implementations

Turnkey Solution (as reference)

Building Blocks & Capabilities to enable 5G Functions & Multi-vertical Edge Services

5G & ORAN
- ONF near-RT RIC + xApps from Intel
- 3rd Party 5GC
- 3rd Party DU/CU using FlexRAN™

OpenVINO™
- Inference Engine (IE)
- OpenVINO ™ Model Server (OVMS)
- OpenCV
- DL Streamer

Intel® Edge Insights for Industrial (EII)

SVT-HEVC for Intel® Xeon®

Intel® CS for WebRTC

Edge Video Analytics Microservice (EVAM)

- Kubernetes Plugins for Intel IPs (eASIC, FPGA, QAT, NICs, GPU in plan for Q4)
- Platform Provisioning
- Security
- Resource Management
- Container Network Interfaces (CNIs)
- Telemetry & Monitoring
- VM Support
- Local/ Distributed Storage

Intel® Smart Edge (built on-top of Kubernetes optimized on IA)

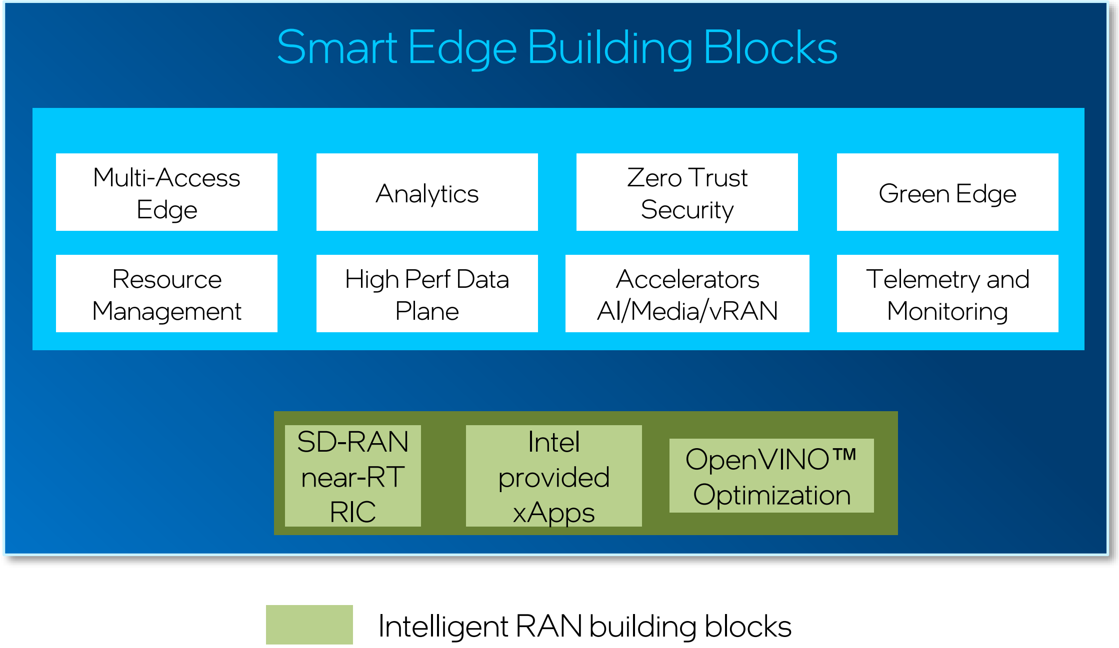
Intel Edge Platform

Intel Recommended HW

Intel® Smart Edge Enabling Intelligent RAN

Smart Edge expands edge AI capabilities to RAN intelligence

Reference solution demonstrating with the SD-RAN near-RT RIC the AI/ML-based Connection Management xApp on 3rd generation Intel® Xeon® platform with significant AI inference optimization through OpenVINO™ to meet real time needs

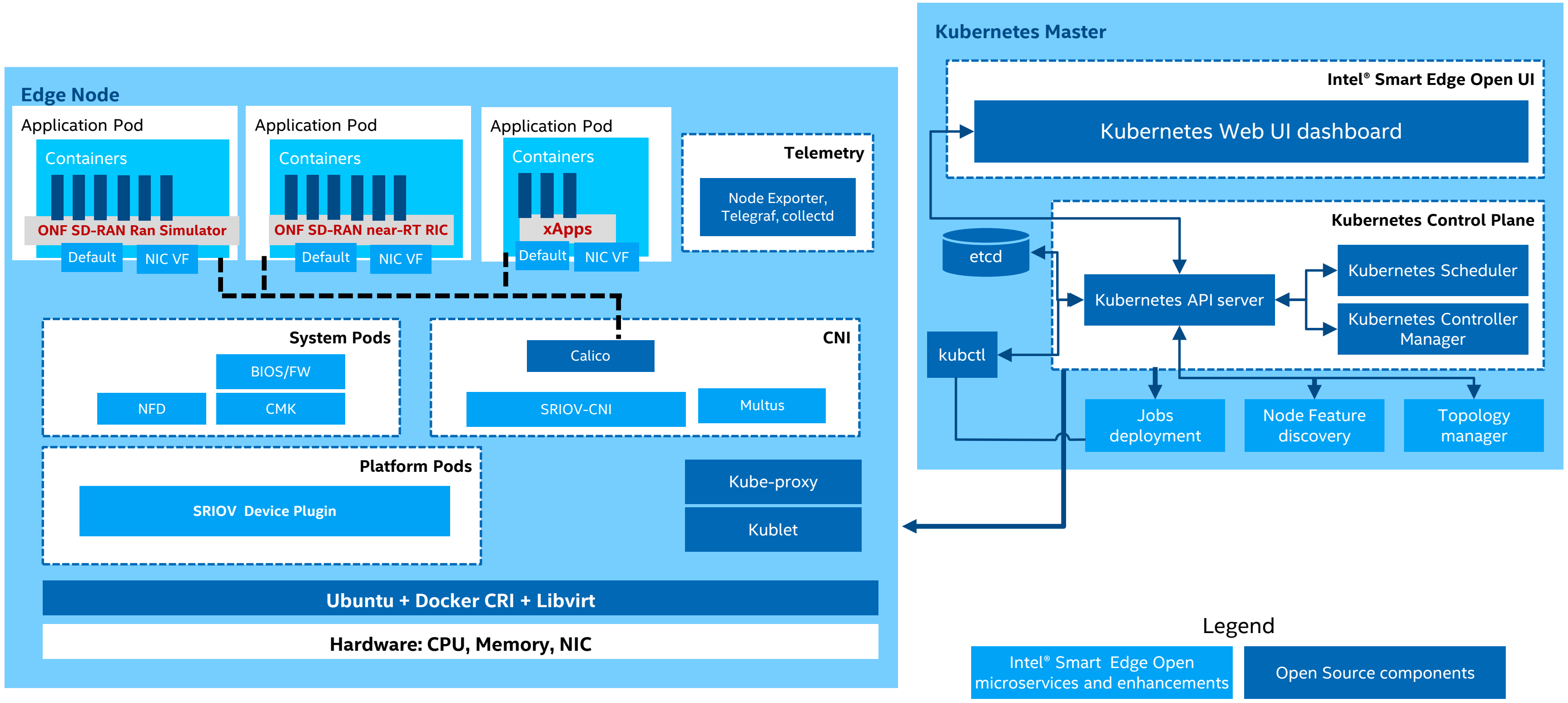


- Reference solution is available through Intel® Developer Catalog

Reference Solution for Near-RT RIC with Intel® Smart Edge Developer Experience Kit

Value-Prop

- Evaluation of RIC & xApps Workload on IA
- Provide xApps from Intel Labs Research
- Validate Performance KPIs for near-RT RIC Use Cases
- Engage with Partners in Tests/Trials Pre-Commercial Deployment
- Jump Start for CoSPs and RIC Vendors Building their own RIC



Key Intel® Smart Edge Building Blocks

Container Orchestration Service	A CNCF certified Kubernetes distribution optimized for CPU, IO and storage resource utilization.
Platform/ Provisioning Service	Configuration and management of low-level platform features in bios and firmware, automated installation of operating system and software stack on bare-metal or virtual machines.
5G RAN/Core Service	Deterministic IO, acceleration, orchestration and platform features for 5G RAN, dynamic re-configuration of RAN and Core
SASE Service	Configuration and management of SASE CNFs enabling threat prevention, web filtering, sandboxing, DNS security, credential theft prevention, data loss prevention and next-generation firewall policies
Zero Trust Security	Platform attestation, integrity, trusted execution, secure device onboarding, and data confidentiality whether data is at rest, in-motion, or in-use
Service Mesh	L3 User Plane and L7 Control Plane service chaining for distributed edge services, with Topology Orchestration UX
Multi-tenancy Service	Resource level isolation and security level isolation implemented using Intel RDT, Kubernetes namespaces and Intel SGX.

Green Edge Service*	Constantly monitors power telemetry and sets power states to reduce the frequency of a CPU core or put a core into a deep sleep state where little or no power is consumed.
Data plane/ Networking Service	High performance data planes and CNIs for various edge use cases: Calico, eBPF, SRIOV, Nodus
Accelerator Service	Kubernetes device plugins for FPGA, FEC, DDP, and GPU*
Storage Service	Persistent storage for stateful edge native apps, with best-in-class read/write throughput and latency on IA, for both single-node and multi-node clusters.
Observability Service	Platform and application telemetry leveraging industry standard frameworks, tuned for minimal core and memory utilization by careful selection of telemetry components.
Video Analytics Service*	Support for Edge Video Analytics Microservice powered by Intel® Distribution of OpenVINO™ toolkit, based on CPU compute or with acceleration.
Operating System Service	OS customization including real time kernel config of kernel threads, interrupt threads, OS services, kernel power services, and housekeeping tasks.

*Coming soon as part of future enhancements

Differentiated Value – Telemetry Building Block

Various platform telemetry could be extracted across management & data plane for differentiated value

Integrated closed loop automation solutions using run time metrics and intelligence

Leverage open-source solutions such as Telegraf, Collectd, Kubernetes Telemetry Aware Scheduler, etc.

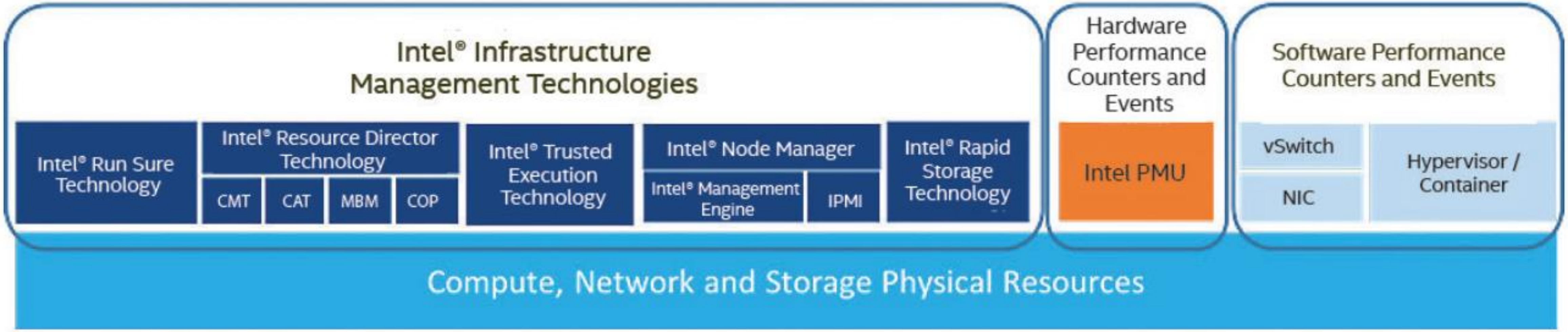
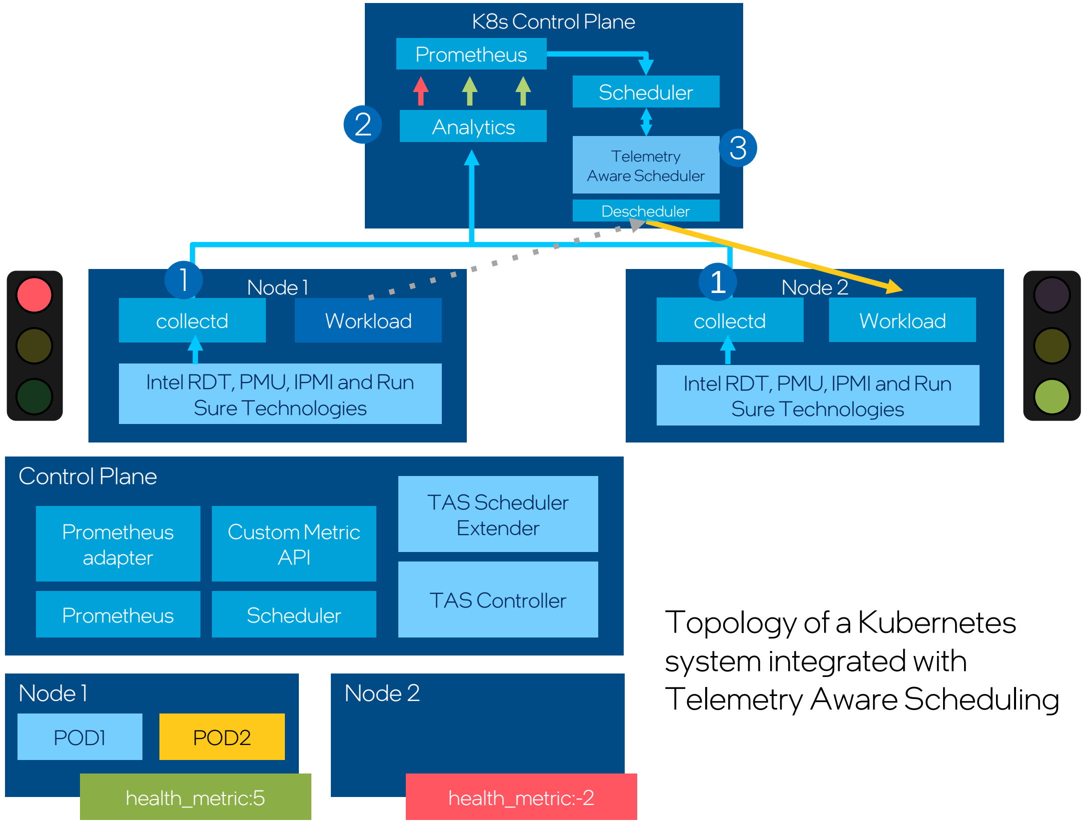


Figure: Platform Telemetry Across the Stack

Resource Management (Telemetry Aware Scheduler)



Topology of a Kubernetes system integrated with Telemetry Aware Scheduling

What does it do?

- TAS collects and exposes platform telemetry from Collectd and other sources to the Control Plane
- Control Plane able to monitor performance of respective nodes and dynamically place/migrate workloads for optimum performance
- Example scenarios:
  - Noisy neighbor situation
  - QoS & QoE fine tuning
  - Platform resilience
  - Real time resource management, etc.

Alternative Solutions:

Absent this level of telemetry integration there is no way to get real time data needed to identify bottleneck and suitable node for dynamic offload

Exposes edge node telemetry metrics enabling service providers to implement rule-based workload placement for optimum performance and resilience

RAN Intelligence Enabling

OCP & Intel® Smart Edge Open

Intel® Network Builders Webinar Aug 23rd 2022

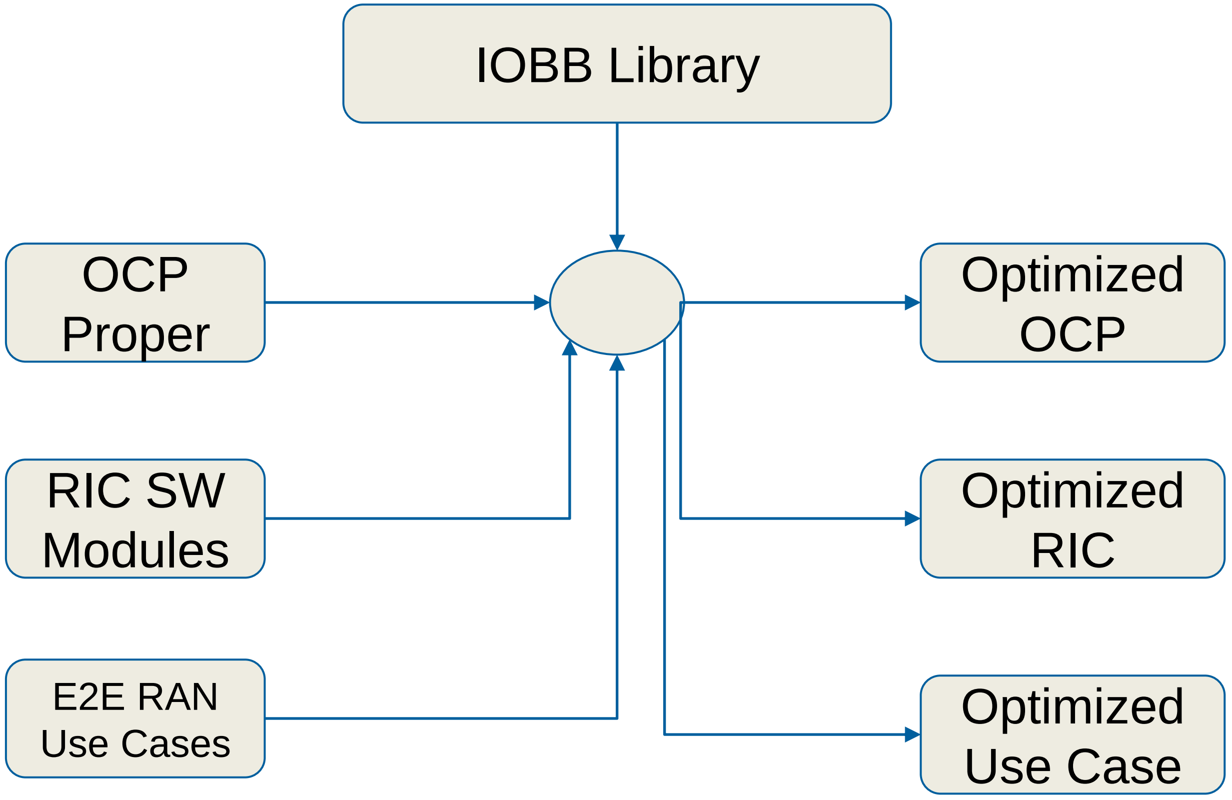
Michael Recchia, Global Telco Solutions Architect, Red Hat
Federico Rossi, Senior Solutions Architect, Red Hat

23 August 2022

Red Hat

Application of Intel Optimization “Building Blocks” (IOBB) for Intelligent RAN

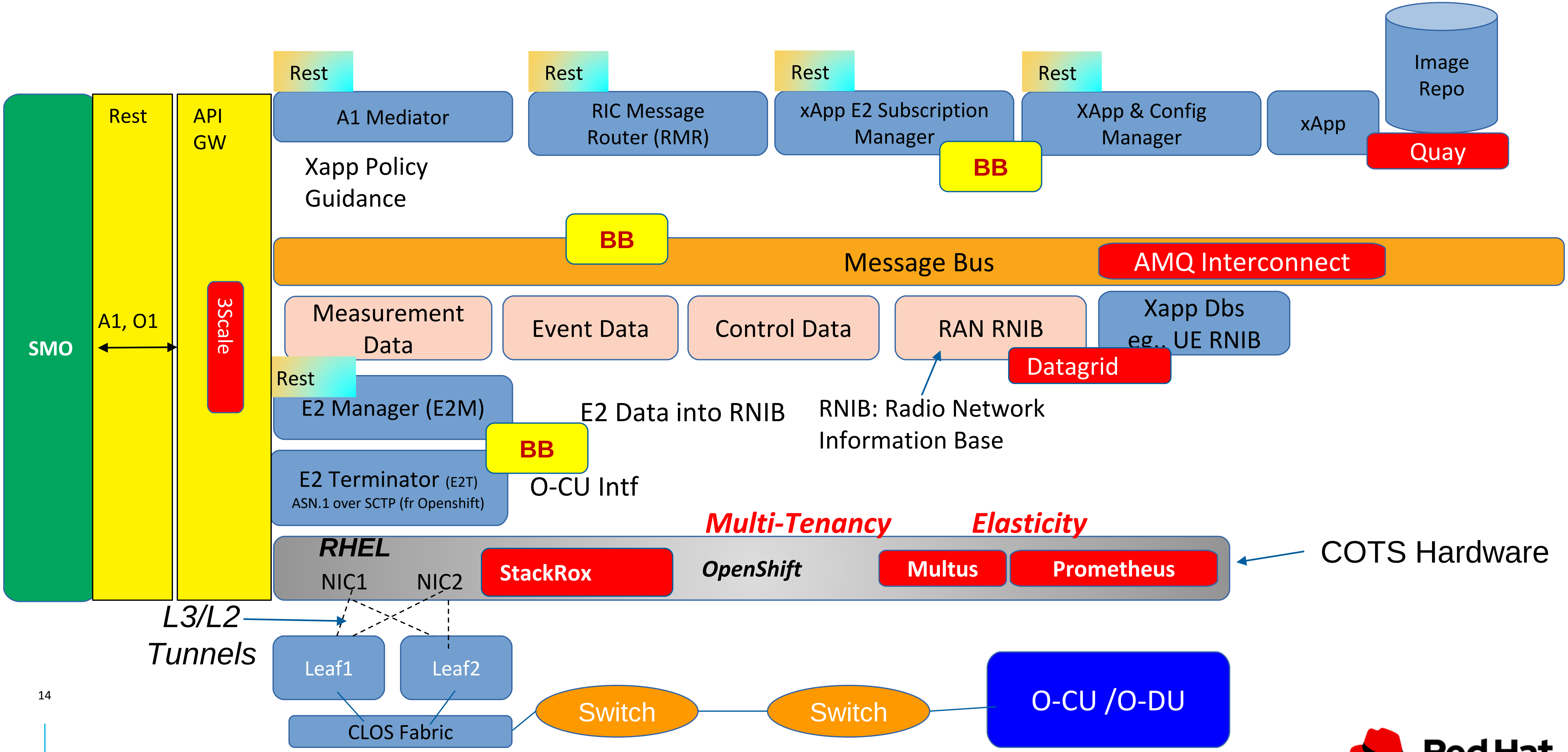
- IOBB are applied to:
  - Process, Application, Resource (PAR)
- PARs are applied within “Categories”
  - OCP Proper (OCP Internals)
    - Includes RHEL OS
  - RIC SW Architecture Components
  - Use Cases
- Output is Optimized PARs within each Category



Categories for this Study

- Openshift Container Platform (OCP) Proper
  - Native (Internal) capabilities of OCP
    - Kubernetes enhancements
    - Networking (SDN), Datapath (e.g., SR-IOV)
    - Operating System
    - HA Capability
    - Virtual Storage
    - Operators, Helm, Service Mesh, Serverless, Pipelines, GitOps, Security, etc
    - Other OCP Internals
- Near-RT RIC Proper
  - Software Modules in the Architecture
    - E2 Termination, E2 Manager, xApp E2 Subscription Manager, etc.
    - RMR (RIC Message Router
    - Data Bus
    - and so on
- Use Case Proper
  - The Network Layer flow (packet flow, packet loss, throughput, latency)
  - Use case objects (SDN, RIC Modules, OCP Modules, Networking Interfaces, etc)

Near-Real Time RIC (Near-RT) Architecture



Red Hat

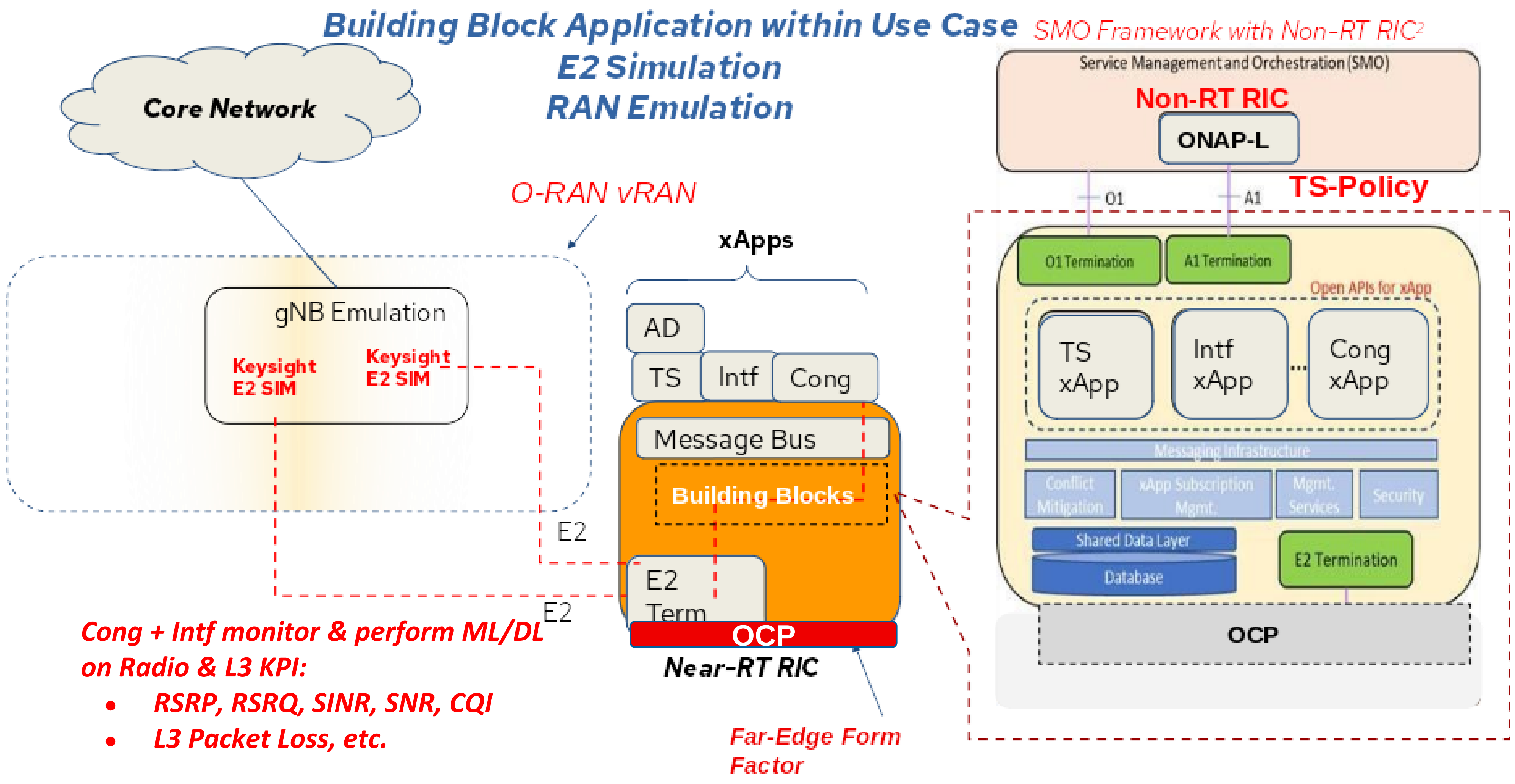
Optimization Vectors
Pathways to Optimization

- Intel Optimization Building Blocks (IOBB) applied to:
    - *RHEL + OCP Services*
    - *Near-RT RIC+xApps, Non-RT RIC+rApps, SMO processes*
    - Red Hat Middleware to optimize Near-RT RIC processes
        - *RIC Message Bus (w/AMQ Interconnect)*
        - *RIC Image Repo (w/Quay)*
        - *RAN RNIB (w/Datagrid)*
        - *E2MGR, E2TERM, RIC Message Router, etc.*
- Cluster workload distribution
    - *N choose M based on "criteria"*
        - *"criteria" = {power consumption, resource loading}*
- *Communicating SLA requirements to the Cloud Layer for E2E Cellular Use Cases, including E2E Network Slicing*

Use Case Definition and Prerequisites

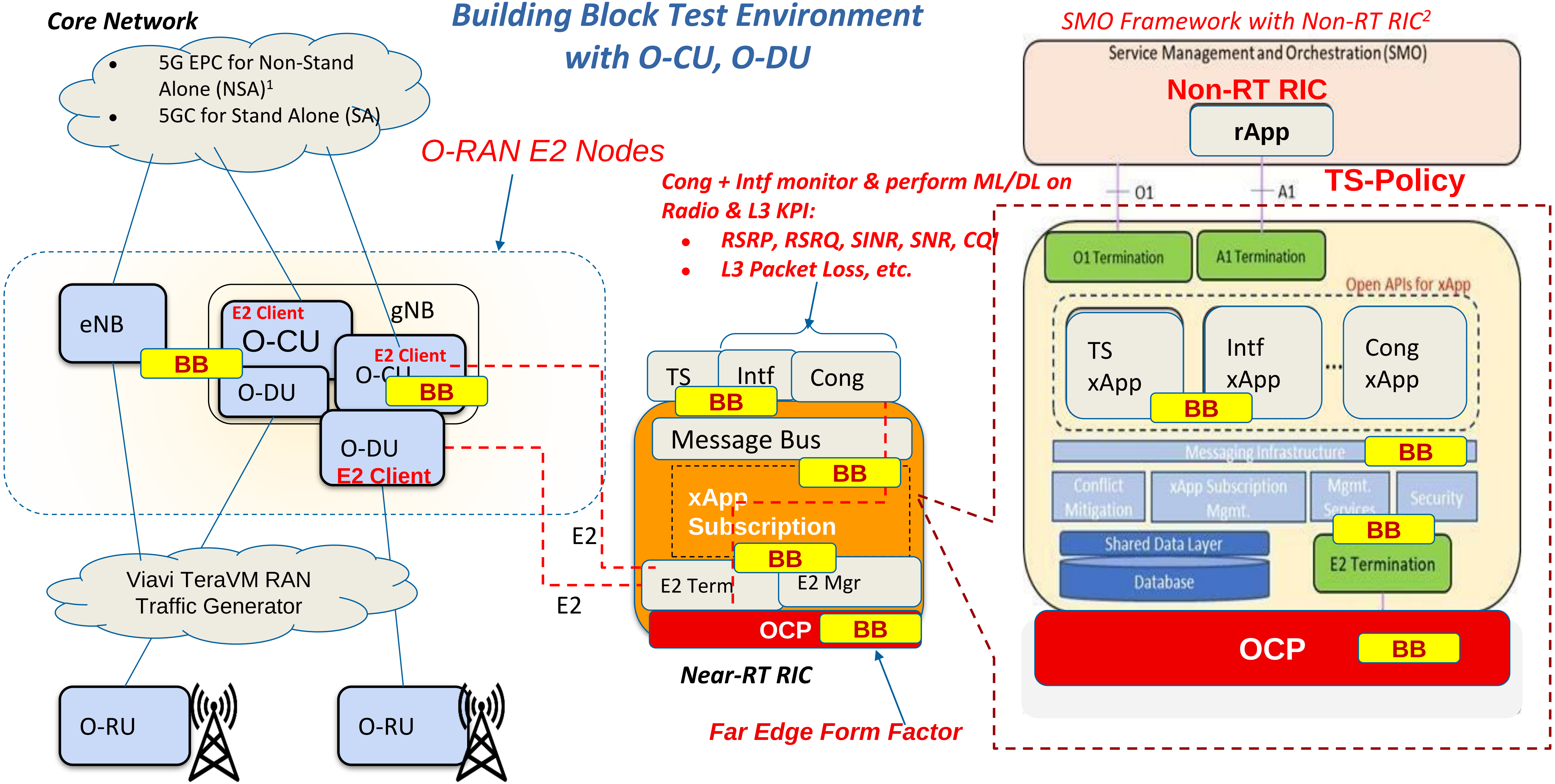
- Use Case Abstract
  - Anomaly Detection with Traffic Steering
  - Start with the O-RAN-SC provided AD, QP, and TS xApps
  - Use new xApps (Congestion/Interference Inferencing) when completed
  - ***Machine Learning (ML) and/or Deep Learning (DL) on Radio & L3 KPI:***
    - ***Congestion & Interference xApps monitor RSRP, RSRQ, SINR, SNR, CQI, L3 Packet Loss, etc.***
      - ***RSRP (Reference Signal Received Power), RSRQ (Reference Signal Received Quality)***
      - ***SINR (Signal to Interference & Noise Ratio)***
    - ***TS xApp then takes action based on ML from the data collected and on A1 policy***
- Prerequisites
  - Far Edge (FE) Foot Print/Form Factor with Intel Chipsets
  - If FE Form Factor is not available
    - Use Dell/HP general purpose servers until FE is available
  - Latest O-RAN-SC Near-RT RIC then other RIC partner
  - OCP 4.10 initially (https://cloud.redhat.com/learn/whats-new)
  - Red Hat Linux, RHEL (Datasheet): https://www.redhat.com/en/engage/redhat-enterprise-linux-s-202109240657?sc_cid=7013a000002w1HhAAI&gclid=EAIaIQobChMI3ePIkbjI9gIVGSiGCh3OMQ3kEAAYASABEgJIh_D_BwE&gclsrc=aw.ds
  - E2 SIM, RAN UP/CP Emulation initially, E2 Nodes/xNb subsequently (O-DU, O-CU)

Initial Lab Topology



Tests: Mobility Patterns (Call Modeling), RF anomalies, RF propagation models (e.g., Path-loss), Fading

Building Block Test Environment with O-CU, O-DU

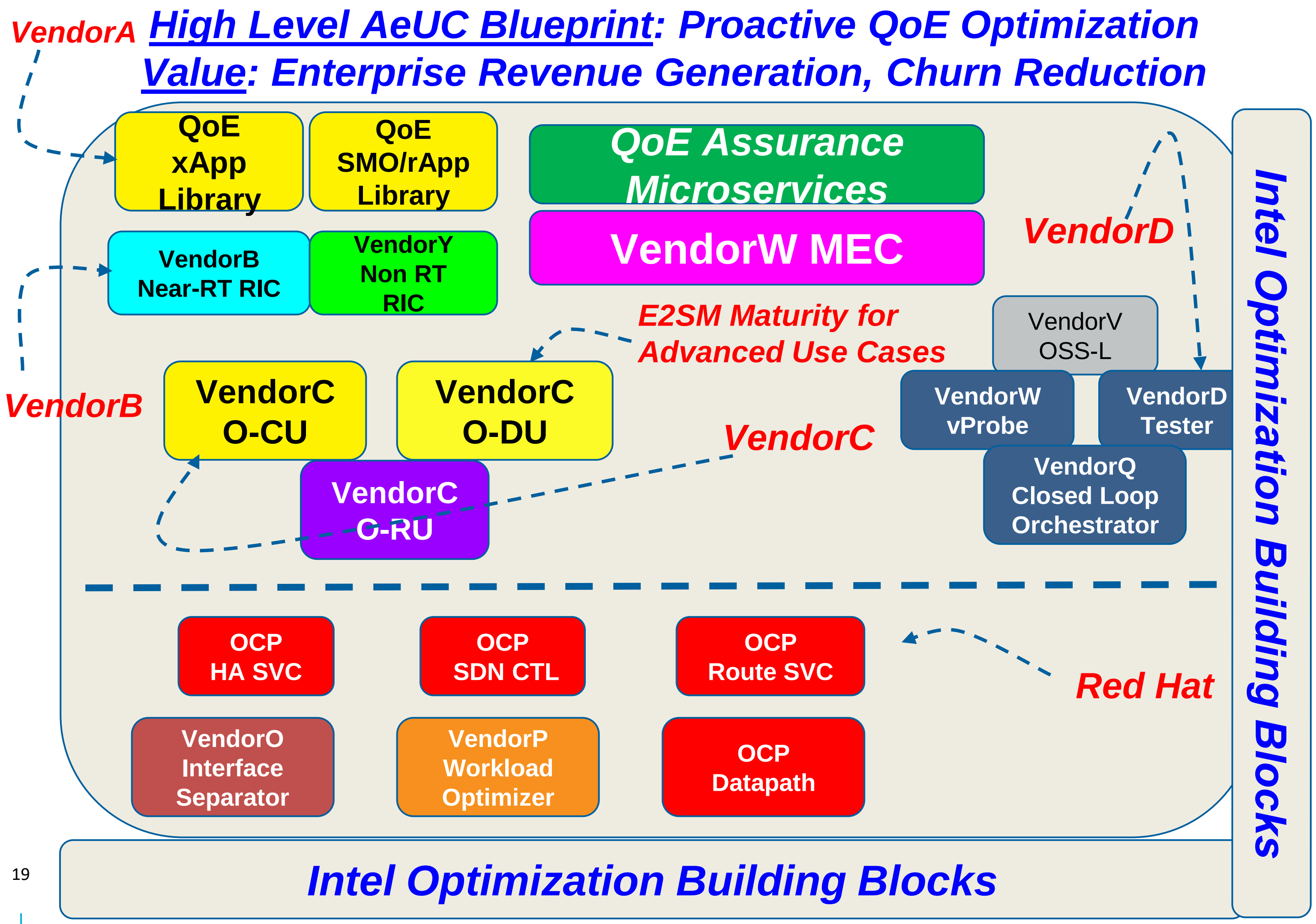


Tests: Mobility Patterns (Call Modeling), RF anomalies, RF propagation models (e.g., Path-loss), Fading

High Level Edge Blueprint for Advanced Intelligent RAN Use Case

Application of Intel Optimization Building Blocks (IOBB)

Telco & Enterprise Services Revenue Growth and OpEx/CapEx Reduction



Application of IOBB for Intelligent RAN

- OS + Cloud + O-RAN Processes
  - RHEL, OCP, O-XU, X-RIC, SMO

Estimate %Time with Max QoE: Three 9s (99.9)

Aggregate Availability: Five 9s (99.999)

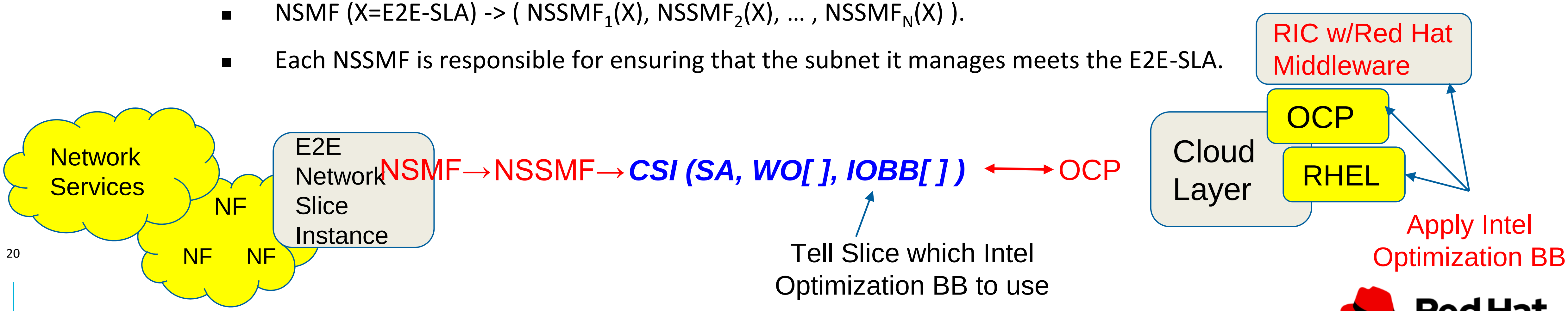
Revenue Estimates (based on partner negotiated contracts):

- Telco customers: $X=Use Case Services Rev
- Private Network customers: $Y=Use Case Services Rev
- xApp Developer: $Rev = %P * Sales
- Near-RT RIC Partner: $Rev = %Q * Sales
- “E2 Client” Partner: $Rev = %R * Sales
- SMO Partner: $Rev = %S * Sales
- MEC Partner: $Rev = %T * Sales
- Non-RT RIC Partner: $Rev = %U * Sales
- Red Hat: $Rev = $Opensource consulting services revenue

Red Hat

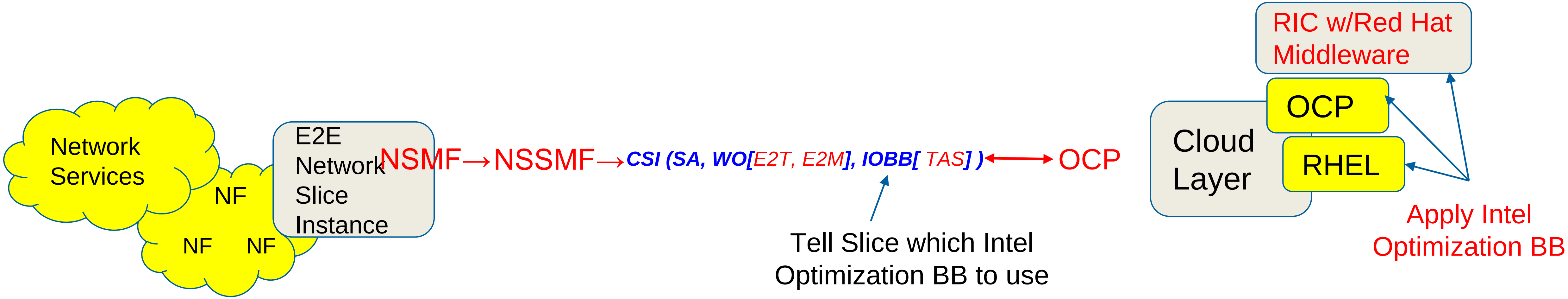
Cloud Signaling with IOBB for E2E SLA Assurance

- Communicate Service Attributes (SA), a Workload Optimization (WO[]) array, and an Intel Optimization Building Blocks array (IOBB[]) from a Network Layer (Application) to the Cloud Layer.
  - Define the E2E SLA & Signal/Pass it to the Cloud Layer
    - E2E-SLA = (SA, WO[], IOBB[]).
      - */* WO[] filled with workload distribution suggestion */*
      - */* IOBB[] filled with Intel Optimization Building Block instructions */*
    - Network Application -> (E2E-SLA) -> Cloud Technology.
  - Enhance Network Slicing via 3GPP CR
    - The Network Slice Management Function (NSMF) passes E2E SLA parameters to each Network Slice Subnet Management Function (NSSMF).
    - NSMF (X=E2E-SLA) -> ($NSSMF_1(X)$, $NSSMF_2(X)$, … , $NSSMF_N(X)$).
    - Each NSSMF is responsible for ensuring that the subnet it manages meets the E2E-SLA.



Red Hat

Example: E2 Optimization with Telemetry Aware Scheduler (TAS)



- *The CSI Workload Optimization (WO) array contains E2 Terminator (E2T) and E2 Manager (E2M). This is in effect telling the Cloud Layer to optimize these 2 processes given the importance of E2 for this use case.*
- *The Intel Optimization Building Blocks (IOBB) contain TAS (Telemetry Aware Scheduler) which is used generally to optimize workload distribution on a Kubernetes Cluster (see next 2 slides).*

OpenShift Telemetry-Aware Scheduling

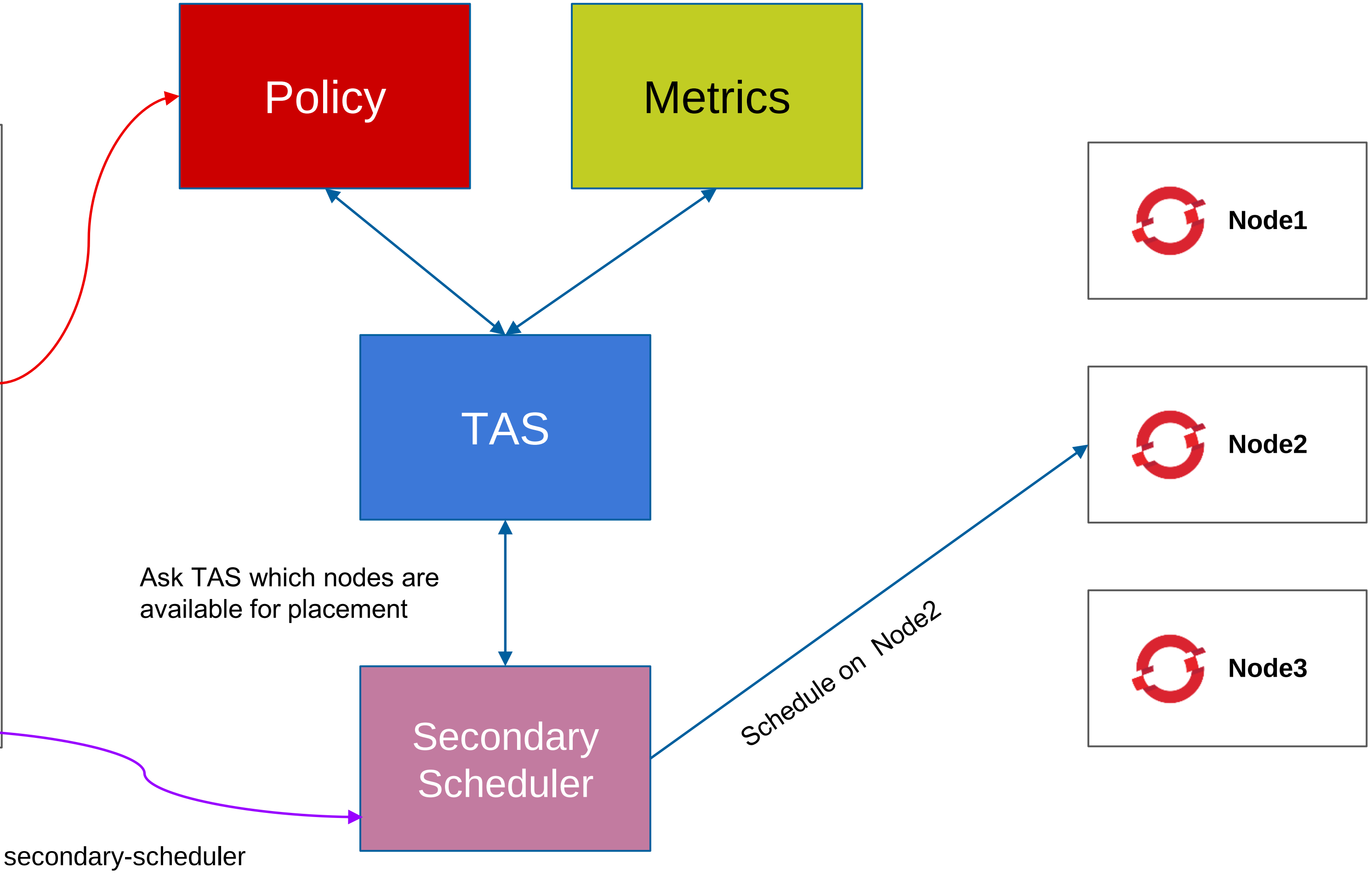
Demo by Federico Rossi, Senior Solutions Architect, Red Hat

Architecture

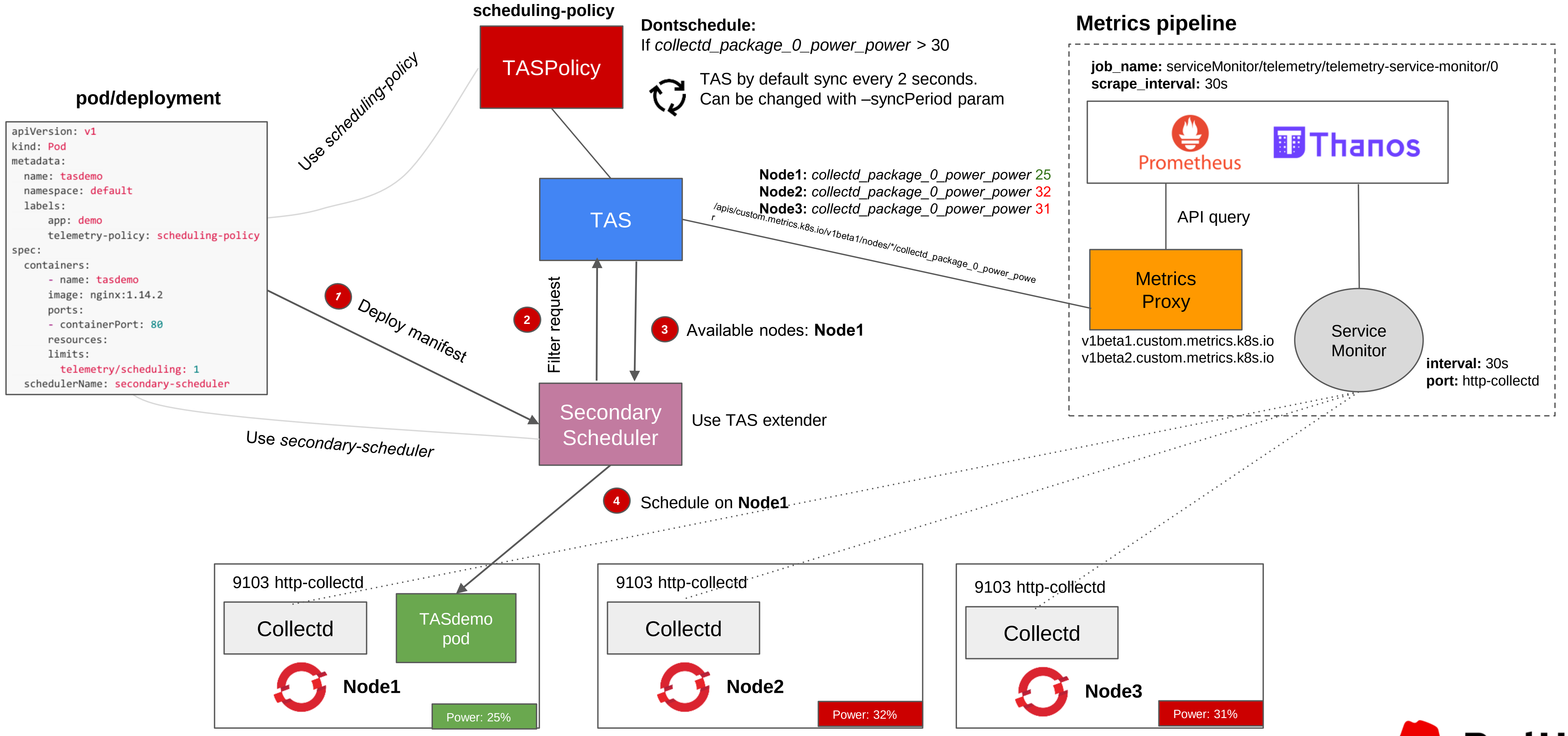
High-level architecture

pod/deployment manifest

```yaml
apiVersion: v1
kind: Pod
metadata:
  name: tasdemo
  namespace: default
  labels:
      app: demo
      telemetry-policy: scheduling-policy
spec:
  containers:
      - name: tasdemo
      image: nginx:1.14.2
      ports:
      - containerPort: 80
      resources:
      limits:
        telemetry/scheduling: 1
  schedulerName: secondary-scheduler
```



Architecture



Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.

linkedin.com/company/red-hat

youtube.com/user/RedHatVideos

facebook.com/redhatinc

twitter.com/RedHat

Red Hat