Model-driven & AI-Enabled Inter-Cloud Optimization

Architecture and Benefits

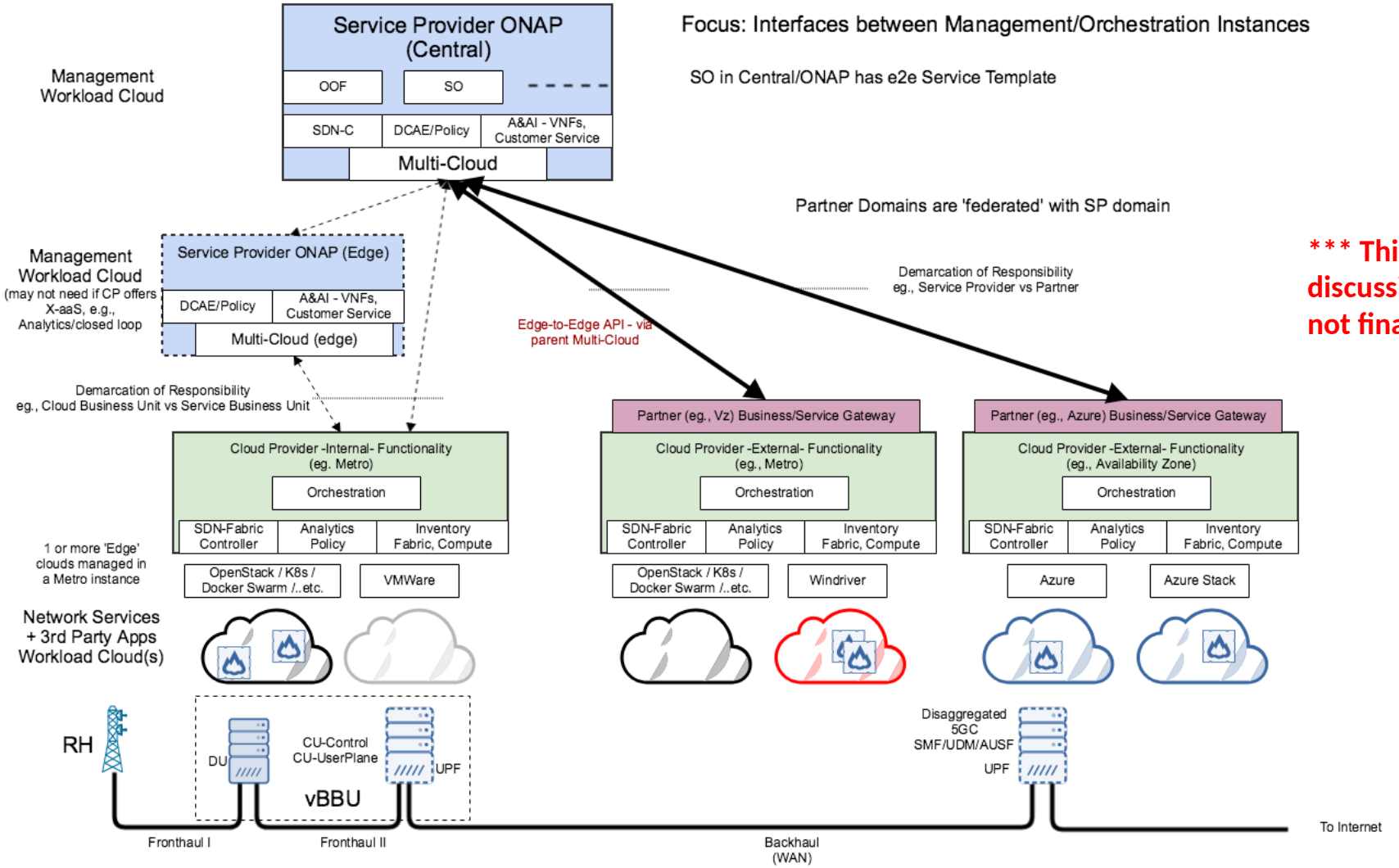
Ramki Krishnan

Introduction

- What did we talk about so far?
  - Model-driven & AI-Enabled Inter-Cloud Optimization
    - 5G/Edge Computing Use Cases – Dilip Krishnaswamy

- Let us talk about the architectural requirements

End-to-end Reference Architecture – ONAP Perspective



*** This diagram is discussion in progress and not final ***

Discussion in Progress: Edge Automation Through ONAP WG (https://wiki.onap.org/display/DW/Edge+Automation+through+ONAP)

Architecture - What do we need? (1)

- Centralized Resource Management/Optimization
  - 1000's of Clouds
  - Probabilistic Decisioning
  - Multiple Solution Choices – Aggregate Data for scale, Data Collection time lag etc.
  - Several Constraints, need flexibility to easily add new constraints
    - Cost (Partner Cloud, Private Cloud etc.), Service SLA (Latency etc.)
  - Data Sources are often Aggregates, examples below
    - Partner/Public Cloud -- Cloud Region & Tenant Resource (Compute/Network/Storage) Available Capacity & Utilization; Cloud Region Energy Utilization
    - Private Cloud – Above + Cluster Capacity/Utilization etc.
  - Policies are often soft constraints, examples below
    - Find Cloud Regions(s) with least resource/energy utilization, least cost etc.
  - Automation Intelligence (AI) through Machine Learning (ML)
    - Use ML (non-linear regression etc.) techniques on operational data to predict the thresholds for soft/hard constraints
    - Update the thresholds for soft/hard constraints in a closed-loop operation

Discussion in Progress: Edge Automation Through ONAP WG (https://wiki.onap.org/display/DW/Edge+Automation+through+ONAP)

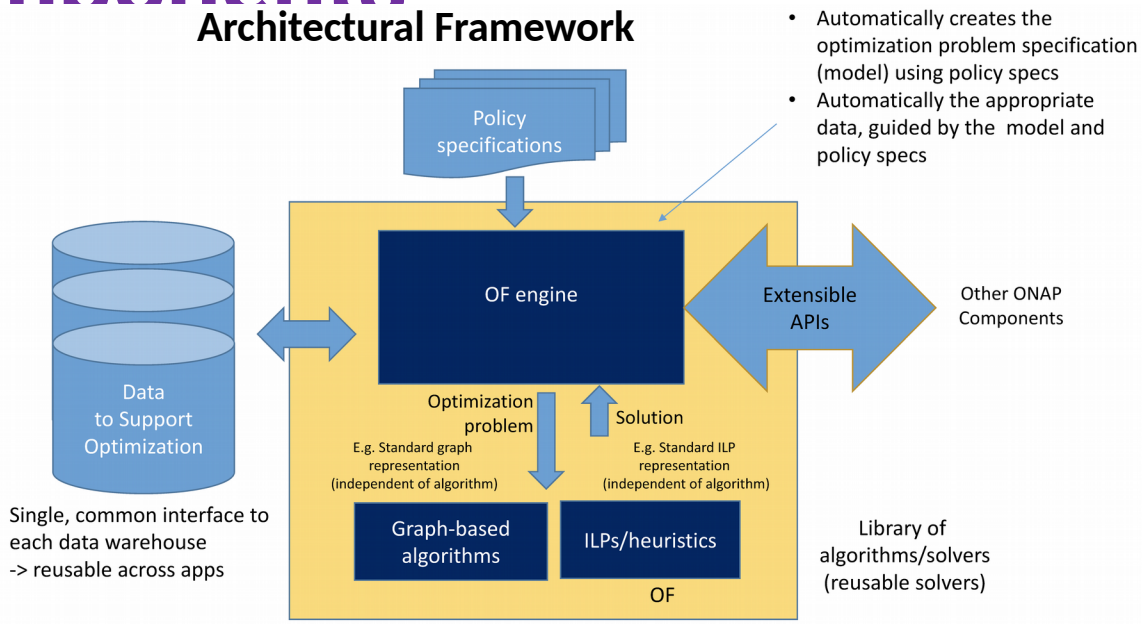
Architecture - What do we need? (2)

- Edge Resource Management/Optimization
  - 1-10 Clouds
  - Accurate Decisioning
  - Single Solution Choice
  - Data Sources are Atomics, examples below
    - Partner/Public Cloud -- Workload (VM/Container) Resource (Compute/Network/Storage) Available Capacity & Utilization etc.
    - Private Cloud – Above + Host Capacity/Utilization etc.
    - Inter-cloud latency, bandwidth etc.
  - Policies are often hard constraints, examples below
    - Find Cloud Regions(s) with SR-IOV support
  - Automation Intelligence (AI) through Machine Learning (ML)
    - Same as Central Resource Management/Optimization
  - Note: For some deployments, this function could be combined with the central component

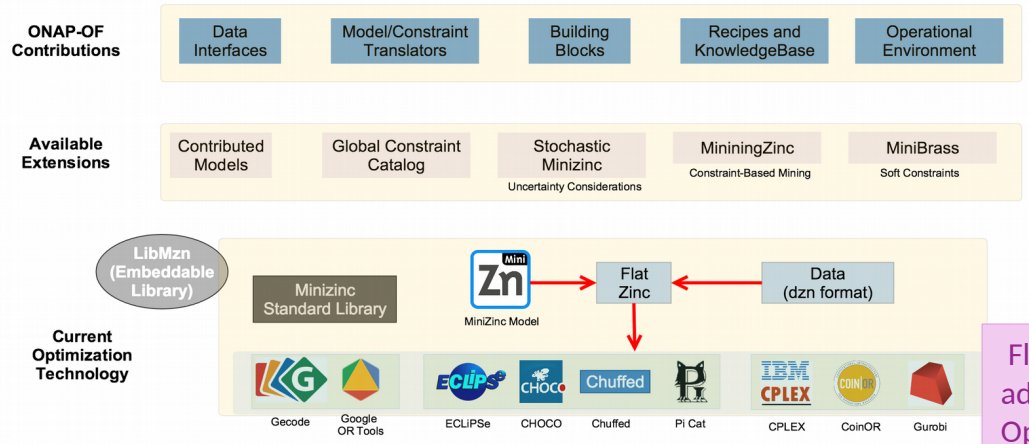
Discussion in Progress: Edge Automation Through ONAP WG (https://wiki.onap.org/display/DW/Edge+Automation+through+ONAP)

Resource Management/Optimization and Related Components

Architectural Framework



Model-driven Optimization Libraries – Minizinc etc.



Flexibility to add Custom Optimizers

ML Component

Use Operational data to predict the thresholds for soft/hard constraints

- **Designer & Developer friendly** Domain-Specific Modelling Language for **Service Placement/Scheduling Policy**
- Address **Central/Edge** Resource Management/Optimization Requirements
- **Masks** the **Mathematical complexity** of optimization algorithms through **Modelling**
- **Flexibility** to add **Custom optimizers** especially for Edge Resource Management/Optimization
- Drive **Service Creation Agility** for 5G, Edge Computing etc.

Discussion in Progress: ONAP Optimization Framework (OOF) -- https://wiki.onap.org/pages/viewpage.action?pageId=3247288

Note: This is an exemplary architectural framework/implementation choice

Upcoming Talks

- “Recent Trends in Constraint Optimization and Satisfaction” -- Nina Narodytska

- “SCOR: Software-defined Constraint Optimal Routing platform for SDN” – Siamak Layeghy
  - Model-driven Minizinc application for constrained-based Routing