AI CLUSTERS

Scaling AI Infrastructure: Meta's Journey with BAG

Jalpa Patel *Technical Program Manager, Meta*

Hany Morsy *Network Engineer, Meta*

OCP PLATINUM Member

Agenda

01

Backend Aggregation (BAG) Overview

Purpose, architecture, and performance benefits

02

Prometheus Case Study

Cluster scaling, topology, and connectivity

03

Engineering Challenges & Solutions

Flexibility, platform fit, distance management, reliability, and capacity planning

Backend Aggregation

Why it exists

- To interconnect L2 fabrics and enable oversized AI workloads
- Without BAG, a cluster can only be as big as one fabric can reach

What it is

- Centralized L3-routed super-spine
- Standard Ethernet end-to-end, no proprietary interconnect

Scale & Performance

- Yields one large cluster across on-prem and leased sites
- Right-sized for the bandwidth needs of inter-L2 workloads



Prometheus Intro

1-Gigawatt Cluster

A massive cluster in New Albany, OH spanning across multiple data center sub-regions.

Innovative Scaling

Accomplished scale with creative deployment solutions:

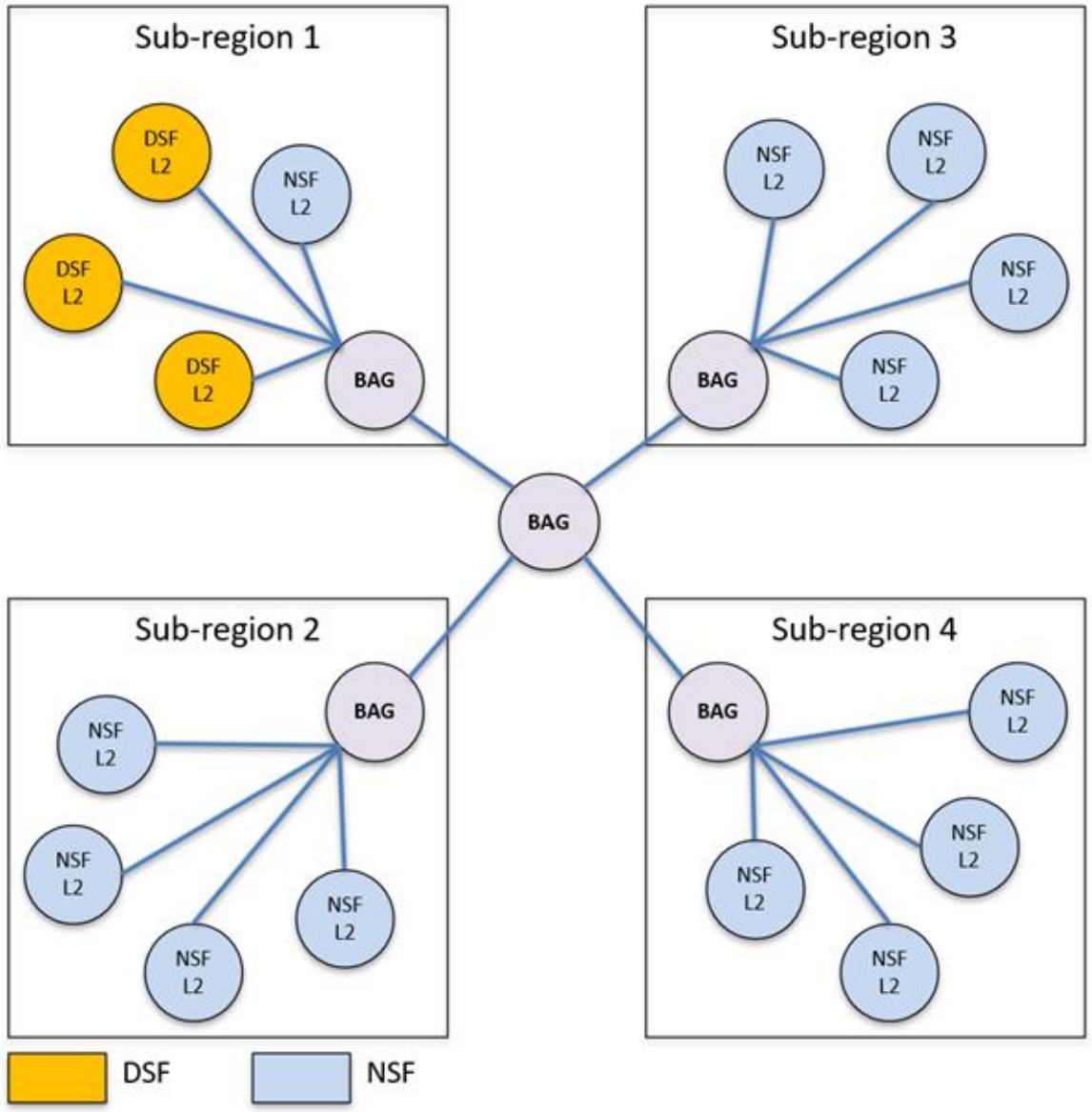
- Traditional data center buildings
- Weatherproof outdoor tents
- Adjacent colocation facilities



BAG in Prometheus

Intra & Inter-Subregion

- Each sub-region holds several L2 fabrics
- L2's attach to their local BAG layer within the sub-region
- All BAG Layers are joined into a single cluster
- BAG is fabric-agnostic. Both DSF and NSF coexist behind it
- Cell-scheduled and packet-sprayed fabrics interoperate through the same L3 layer



Zoom-in to the NSF Topology

Scale & Capacity

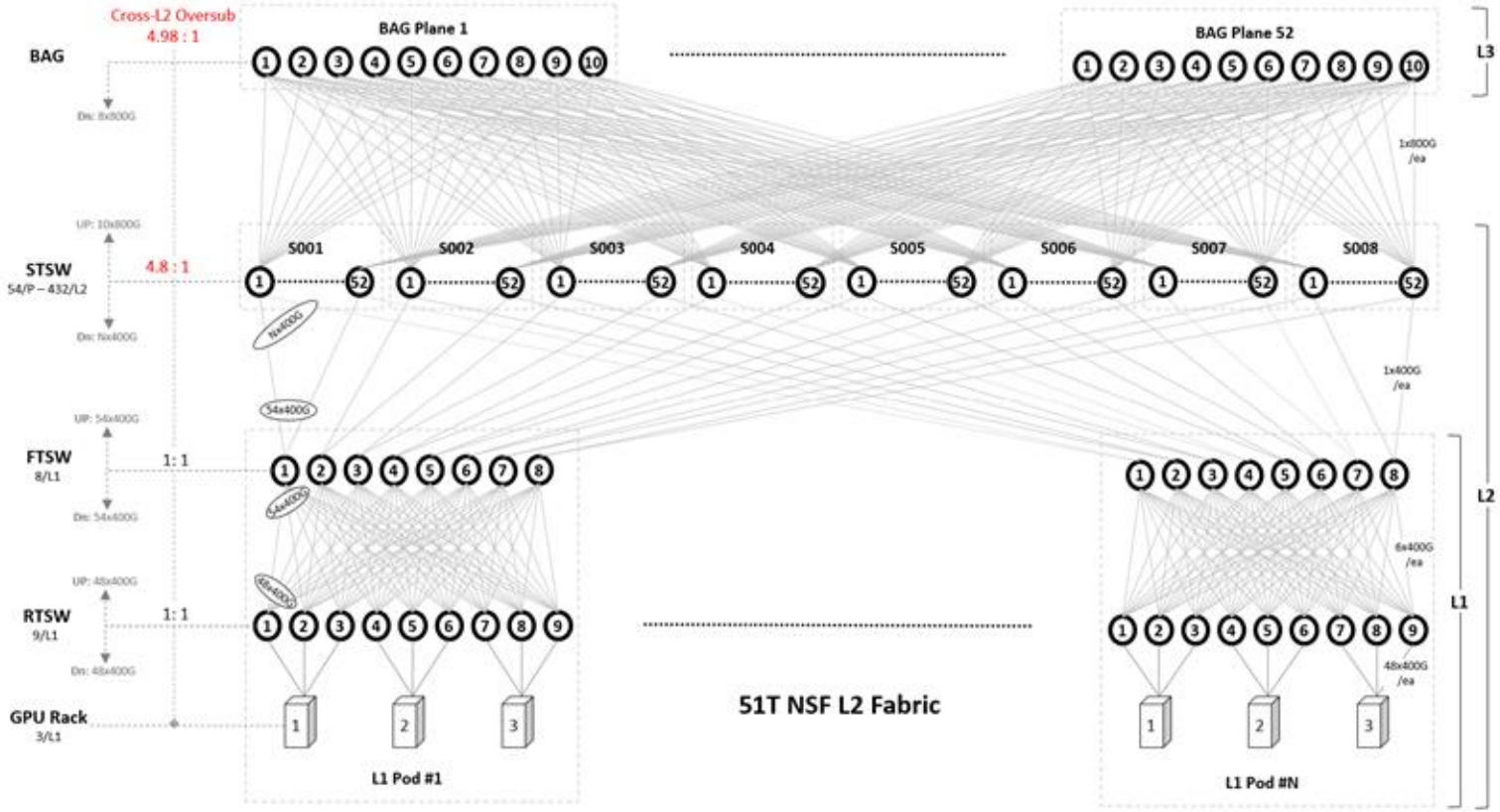
- Up to ~21K GPUs per L2 fabric
- Intra-Pod and Inter-Pod both are **Non-Blocking**

Backend (BE) Pod Design

- **Each Pod:** 9 RTSWs + 8 FTSWs + 3 GPU
- Non-rail, SoR rack
- Each GPU rack serviced by **3 RTSWs**
- GPU rack can be liquid cooled or AALC

Switch Planes & Oversubscription

- **8 Spine Planes** (54 STSWs/Plane, Intra
- **~5 : 1** Cross-L2 Oversubscription



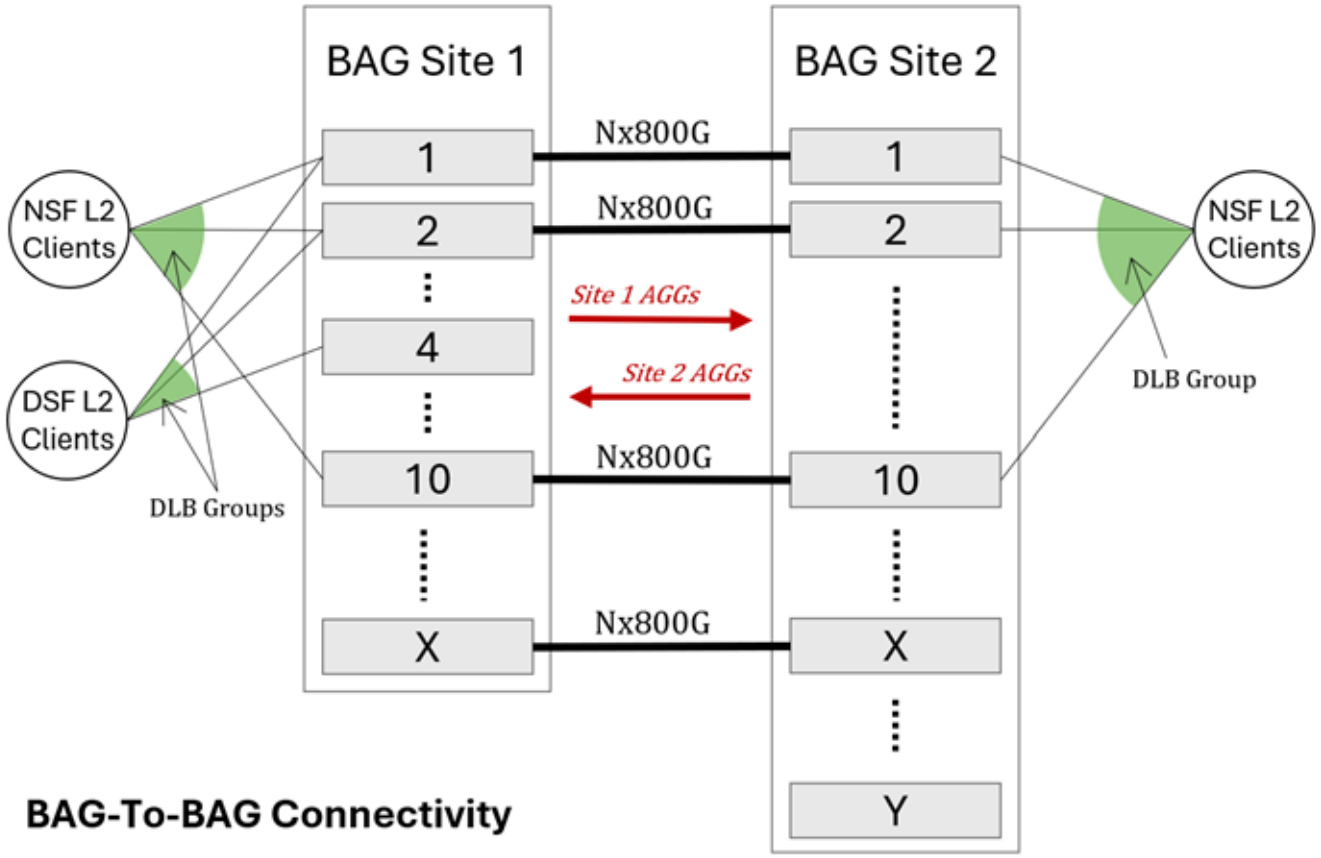
BAG-to-BAG Connectivity Scheme

Planar Connectivity

- BAG connects only to BAG_i across sites — N × 800G per pair
- Simpler cabling with independent parallel planes, not a mesh
- Unmatched BAGs at the larger site carry local L2 traffic only

Forwarding

- L2 aggregates propagate site-to-site through the BAG
- Each STSW forms a DLB group across its BAG plane to reach remote L2s
- RTSWs spray load-aware DLB end to end — intra-L1 through inter-BAG

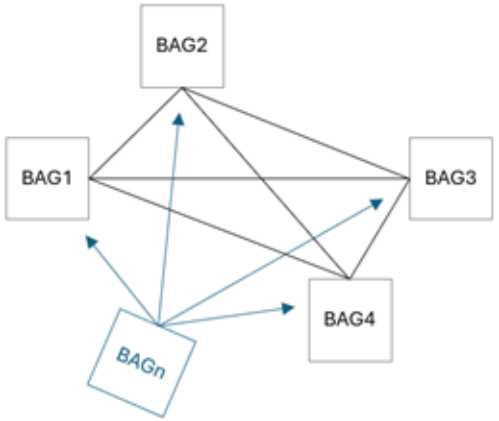


BAG-To-BAG Connectivity

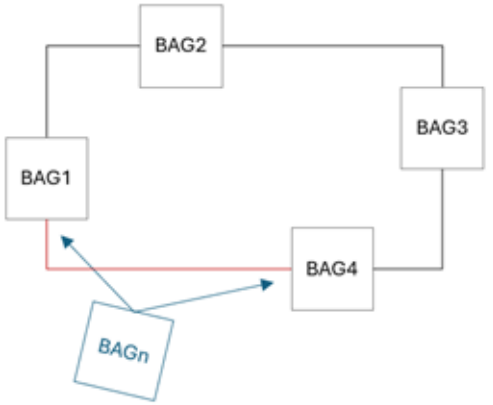
BAG Challenges - Flexibility

Adding new sites flexibly

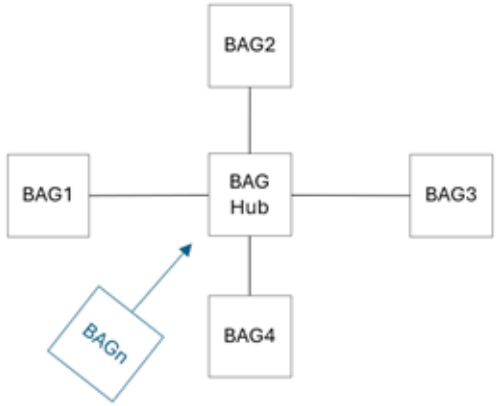
- **Challenge:** Adding new sites in Prometheus on-demand and connecting it to the BAG Backbone.
- **Solution:** Build the Bag-to-Bag topology as Hub & Spoke
- Full mesh: adding a site touches every existing site — N new circuits
- Ring: only 2 touch points, but a live link must be replaced
- Hub & spoke:
  - Our preferred choice in Prometheus
  - One touch point — the hub
  - Links grow O(N), not $O(N^2)$
  - spokes hold no radix for sites that don't exist yet
  - We deploy a larger chassis form-factor in the Hub



Full Mesh



Ring



Hub & Spoke

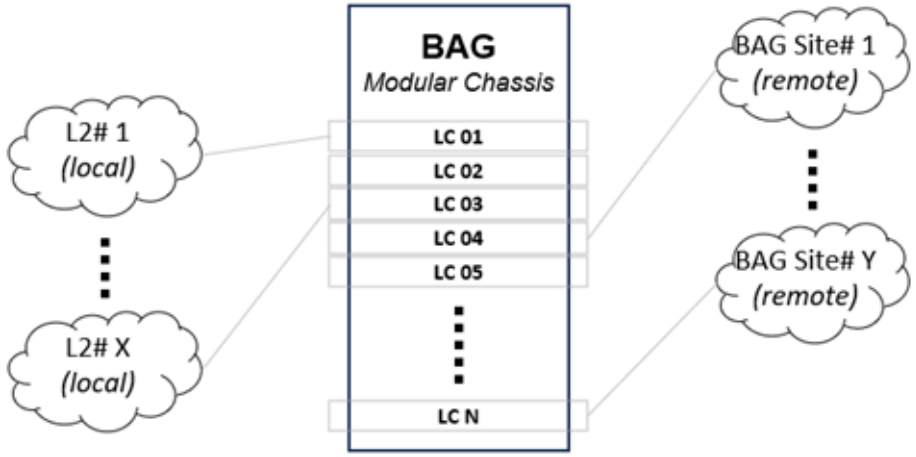
BAG Challenges – Platform Fit

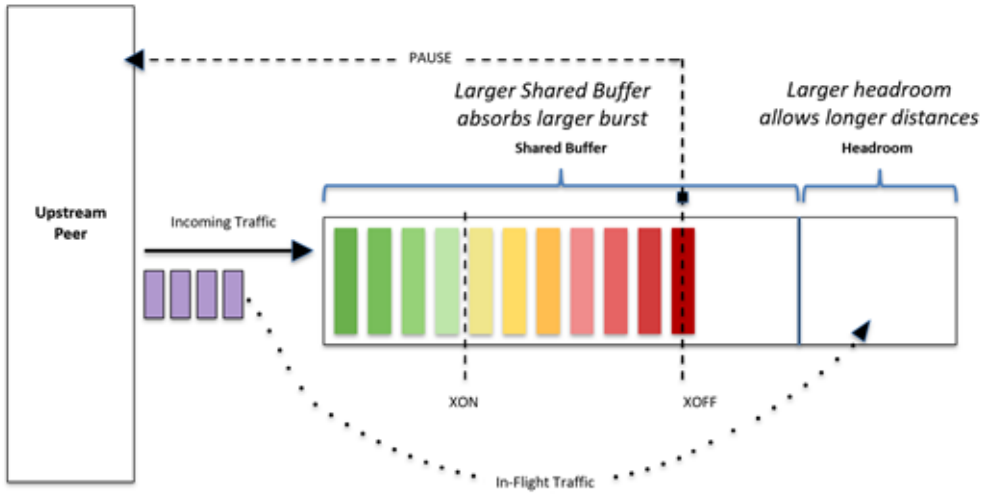
Radix

- Challenge: Terminating many local L2's and several remote BAG sites.
- Solution: Select a Modular Chassis with large total port radix

Long Haul RoCEv2

- Challenge: Ensuring lossless delivery over long haul fiber paths.
- Solution: Choose Deep Buffer ASIC with credit-based forwarding algorithm – J3 Class
  - Larger headroom allows longer distances
  - Larger shared buffer absorbs larger bursts

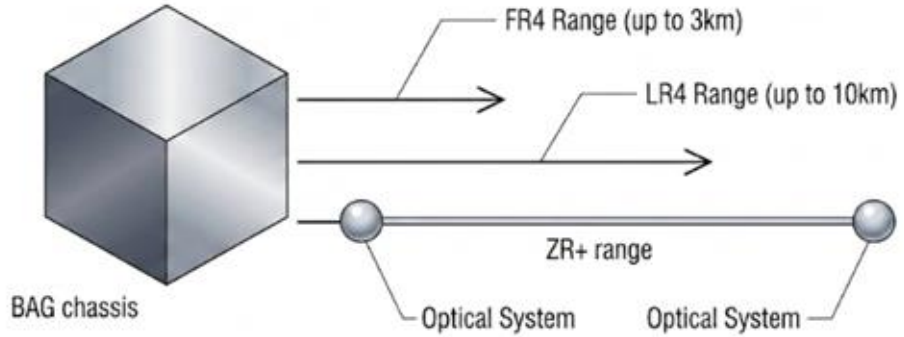




BAG Challenges – Distance Variability

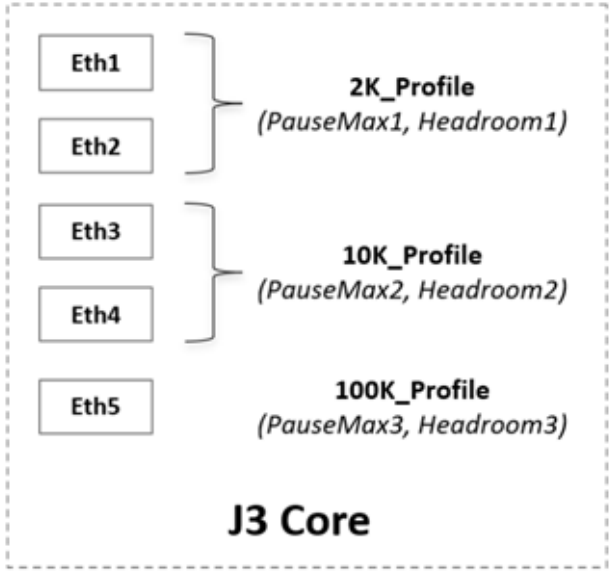
Pluggable Transceiver Mix

- Challenge: BAG terminates a variety of fiber distances
- Solution: Platform support for a mix of optics
  - FR4 Up to 3 km
  - LR4 Up to 10 km
  - ZR+ Up to 1000 km



FADT Multi-Profiles

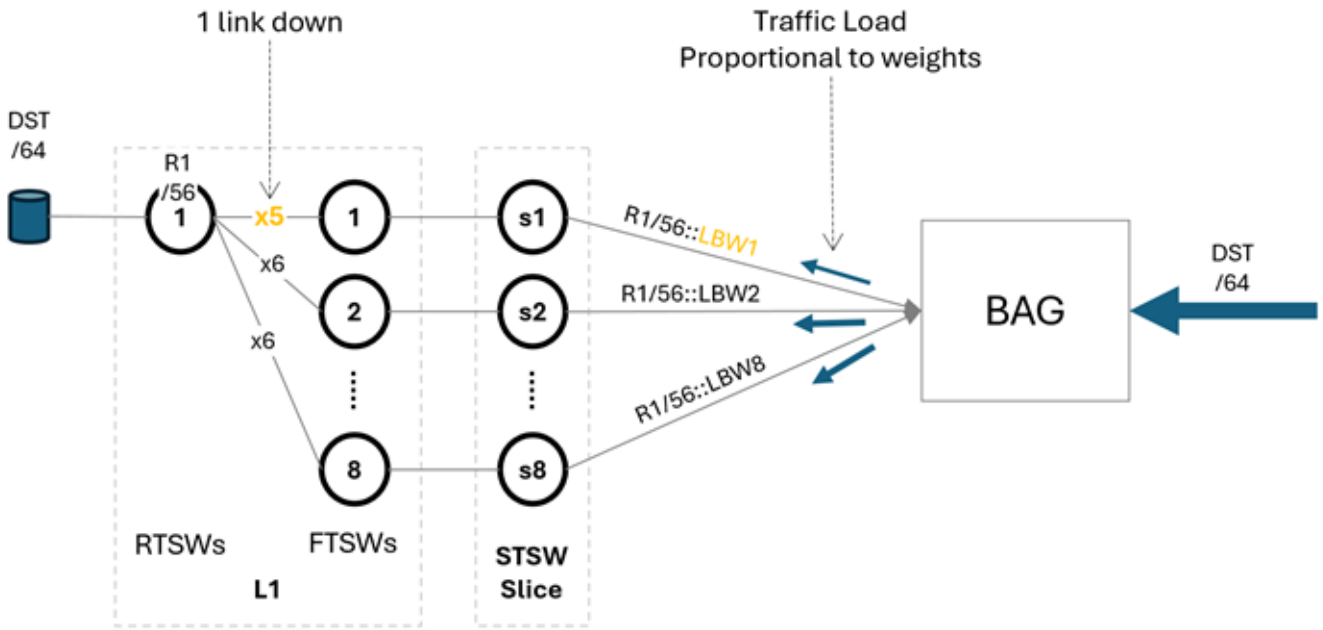
- Challenge: Efficient Buffer carving to fit the length variability
- Solution: Apply FADT Multi-Profiles with variable headroom allocations based on distance



BAG Challenges – Failure Awareness

WECMP

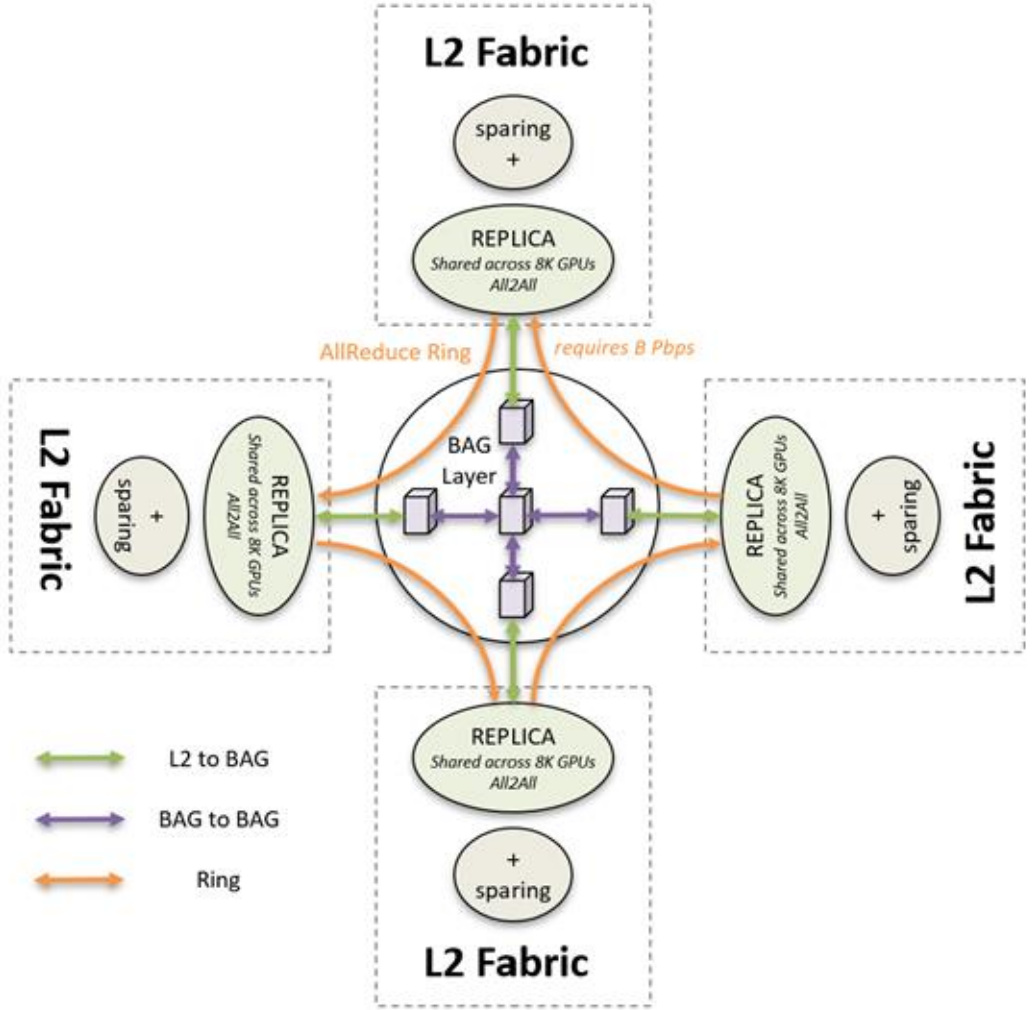
- **Challenge:** ECMP assumes every path has equal capacity. How can the BAG account for failures deep inside the L2 fabrics it is connected to?
- **Solution:** We deploy WECMP on the BAG where it reacts to LBW changes within its fabric and adjusts its load proportions to avoid bottlenecks.



BAG Challenges – Right Sizing BAG Capacity

Driven by Model Replica

- **Challenge:** Determining the right amount of BW for:
  - L2 to BAG
  - BAG-to-BAG
- **Solution:**
  - A Replica is placed fully within the L2, allowing the BW-intense collectives to use the non-blocking network.
  - DP runs across replicas over the BAG using AllReduce Ring. Provides cross-replica sync every training step.
  - Minimum B Pbps required by the AllReduce Ring.
  - Size L2-to-BAG capacity for B + *buffer*. Buffer is for operational maintenances.
  - Size BAG-to-BAG capacity based on availability of optical protection – (B + buffer) x 2 in absence of protection



Call to Action

- Come join us to solve interesting, complex problems at scale in AI infrastructure.
- Help address technical challenges in long-distance training and network buffer management.
- SAI collaboration to enable BAG functions on OCP switches
- Join the OCP Networking Project to help define future open-source infrastructure specifications.
- Reach out to us at jalpatel@meta.com or hany@meta.com to discuss further collaboration.

Thank You!

