

ENABLING AI Infrastructure

**Tomahawk Ultra – Ultra Low Latency,
High Bandwidth Ethernet Switch for
HPC & AI/ML applications**

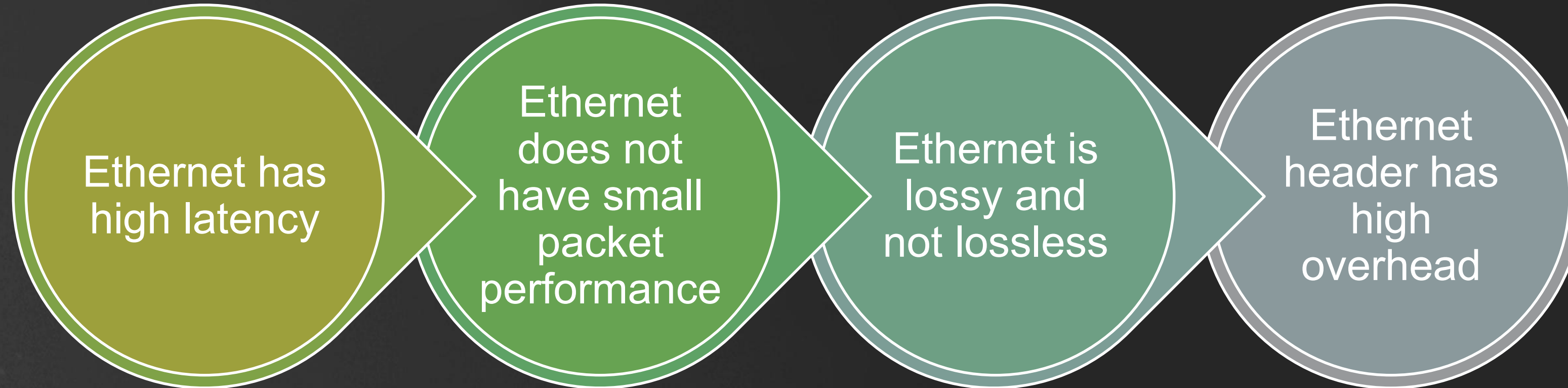
Mohan Kalkunte & Asad Khamisy

August 25, 2025

HPC and AI Fabric Demands:

Attribute	HPC	AI Scale-Up
Small Message Transfer	✓✓	✓
High Fabric Throughput	✓✓	✓✓
High Bandwidth	✓	✓✓
Ultra Low Latency	✓	✓
Low Protocol Overhead	✓	✓
Deterministic Transfer	✓	✓
Scalable	✓✓	✓
Topologies	Clos, DragonFly, Mesh, Torus	Clos, Mesh, Torus

Ethernet Misconceptions



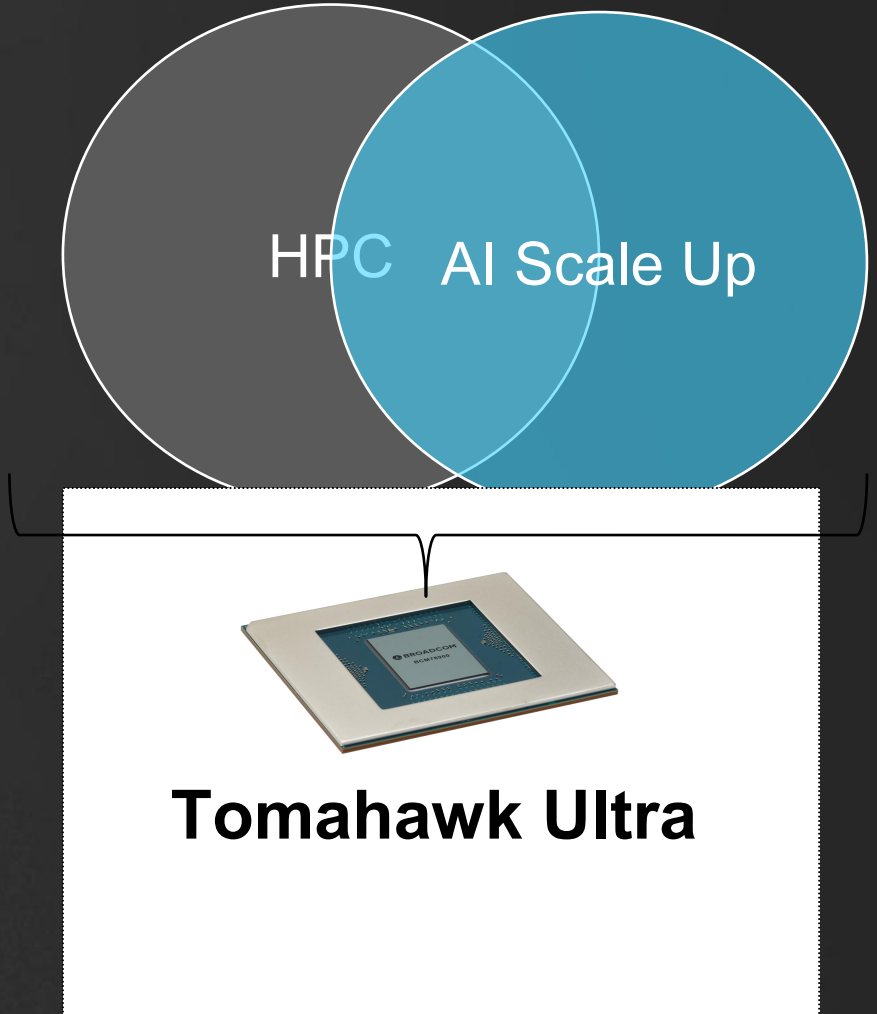
Introducing Tomahawk Ultra

NOW SHIPPING



Tomahawk Ultra
512 x 100G-PAM4

Tomahawk Ultra Highlights – Covering HPC and AI Scale Up



Performance

250ns latency
51.2T@64B
77 BPPS



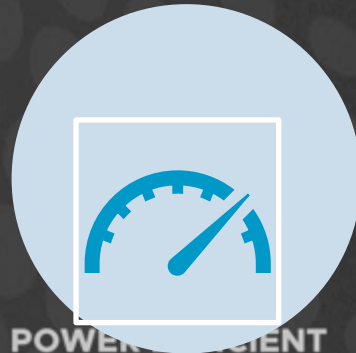
Efficiency

In-Network Collectives
Optimized Ethernet Headers



Reliability/Lossless

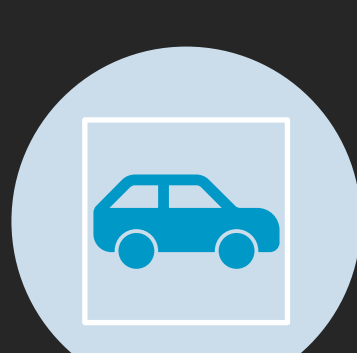
LLR
CBFC



Programmable Visibility/Telemetry



Topology Aware Routing

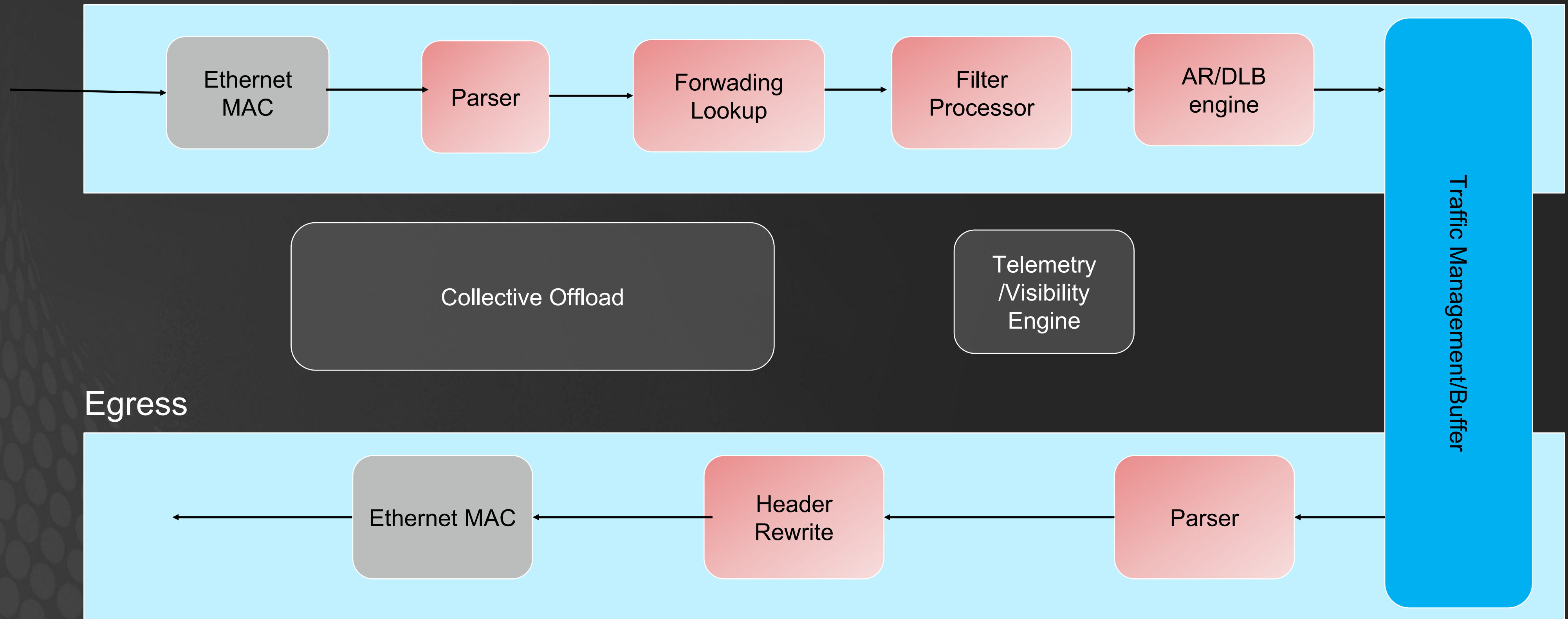


Congestion Control



Packet Forwarding Pipeline

Ingress



Egress

Key Features described

LLR & CBFC

Optimized
Header

In-network
Collectives

Topology
Aware
Routing

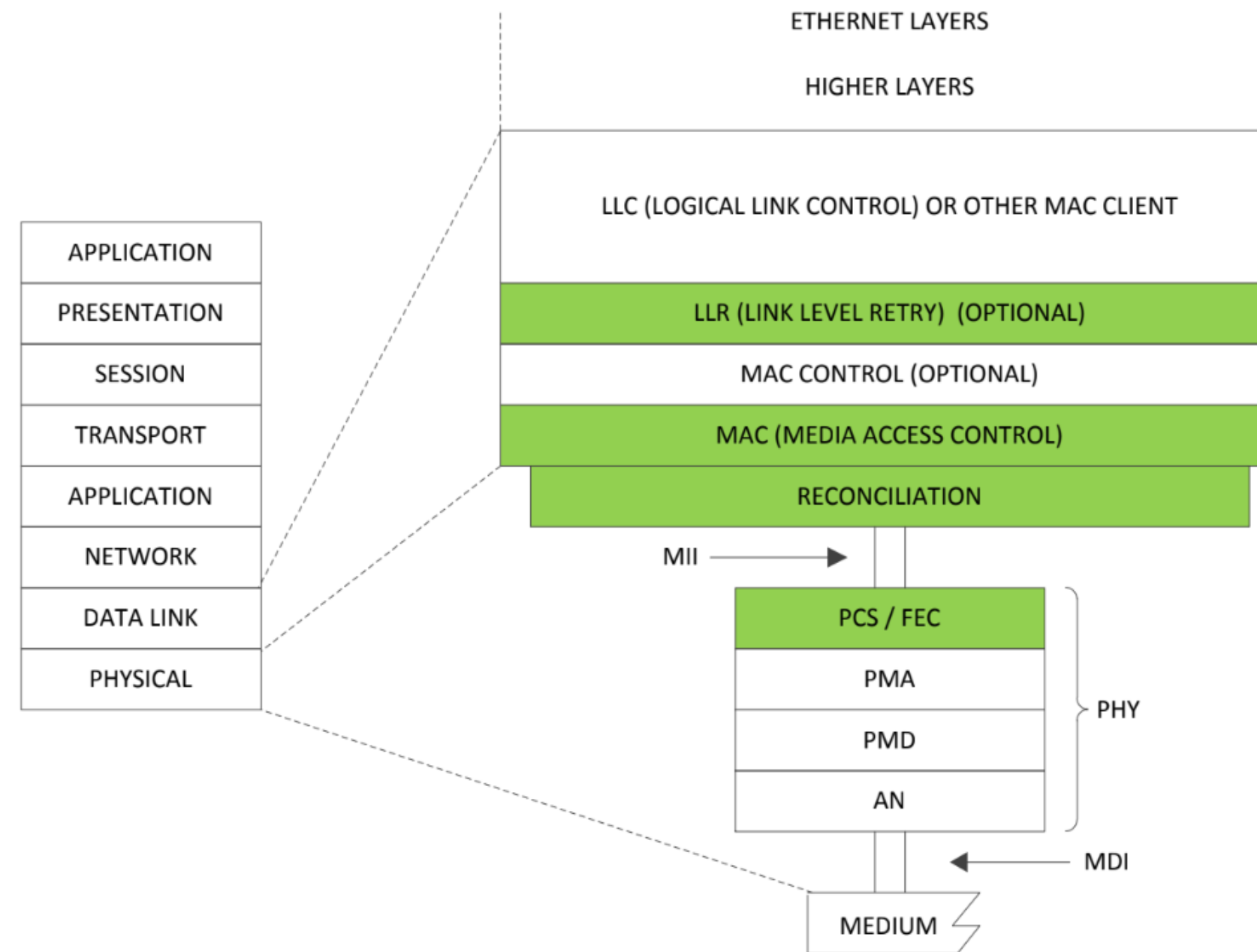
Congestion
Control

Programmable
Visibility

Performance

Link Layer Retry

Architectural Overview



Benefits

- ❖ LLR complements the Ethernet FEC.
- ❖ Improves link robustness for bursty errors or sub-optimal links.
- ❖ Reduces the need for high latency end-end retransmissions.

Link Layer Retry



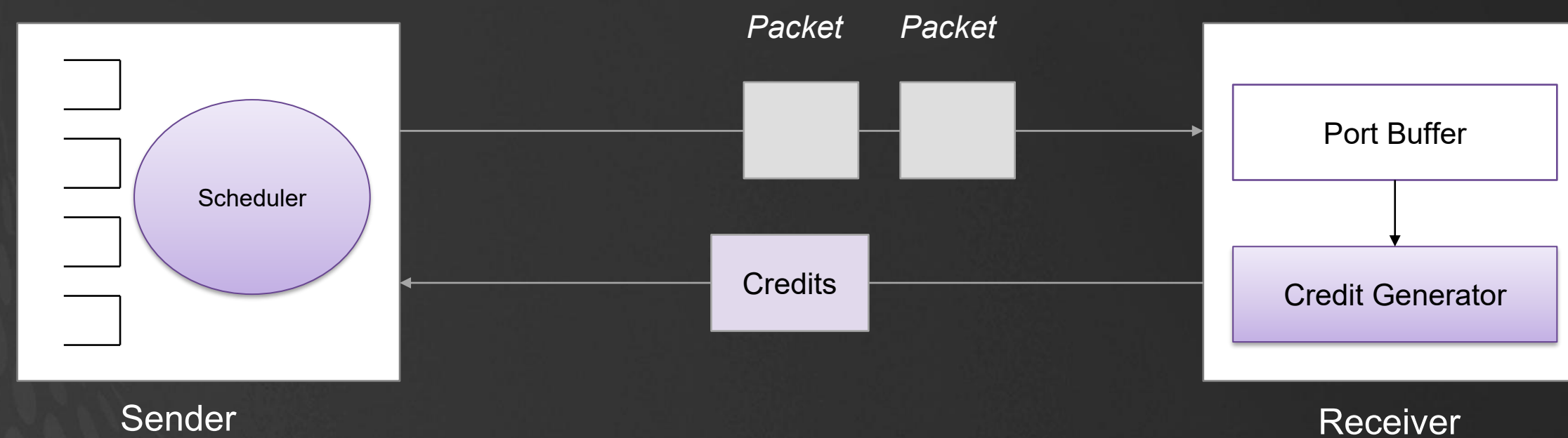
Transmitter

- ❖ Ethernet Preamble to carry a packet's sequence number
- ❖ Replay buffer stores transmitted packets until ACK reception
- ❖ Timeout mechanism

Receiver

- ❖ Tracks expected sequence number.
- ❖ Updates it when it receives an error-free packet in the correct sequence.
- ❖ Sends ACK for good packets, NACKs for corrupt or out-of-sequence packets.
- ❖ Filters packets that don't match expected sequence number.

Credit Based Flow Control (CBFC)



- **Receiver**

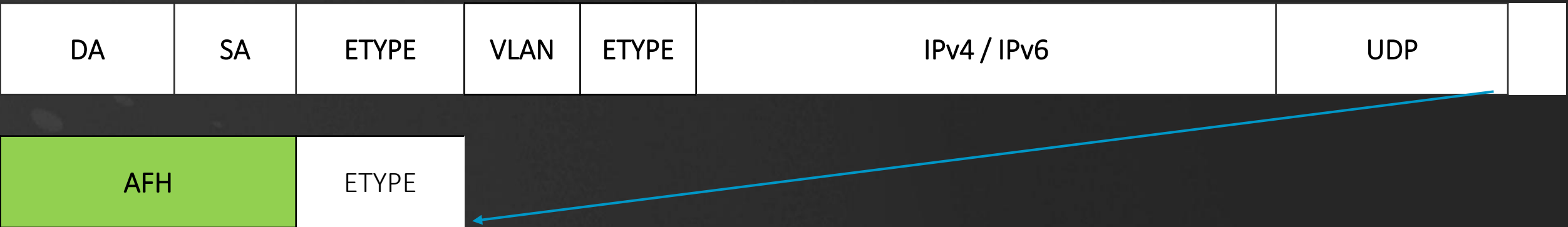
- ❖ Reserves buffer for each lossless traffic class.
- ❖ Reserved buffer can be dedicated or shared across different traffic classes.
- ❖ Sends credits to transmitter based on buffer availability.
- ❖ Stops sending credits when buffer is full.

- **Sender**

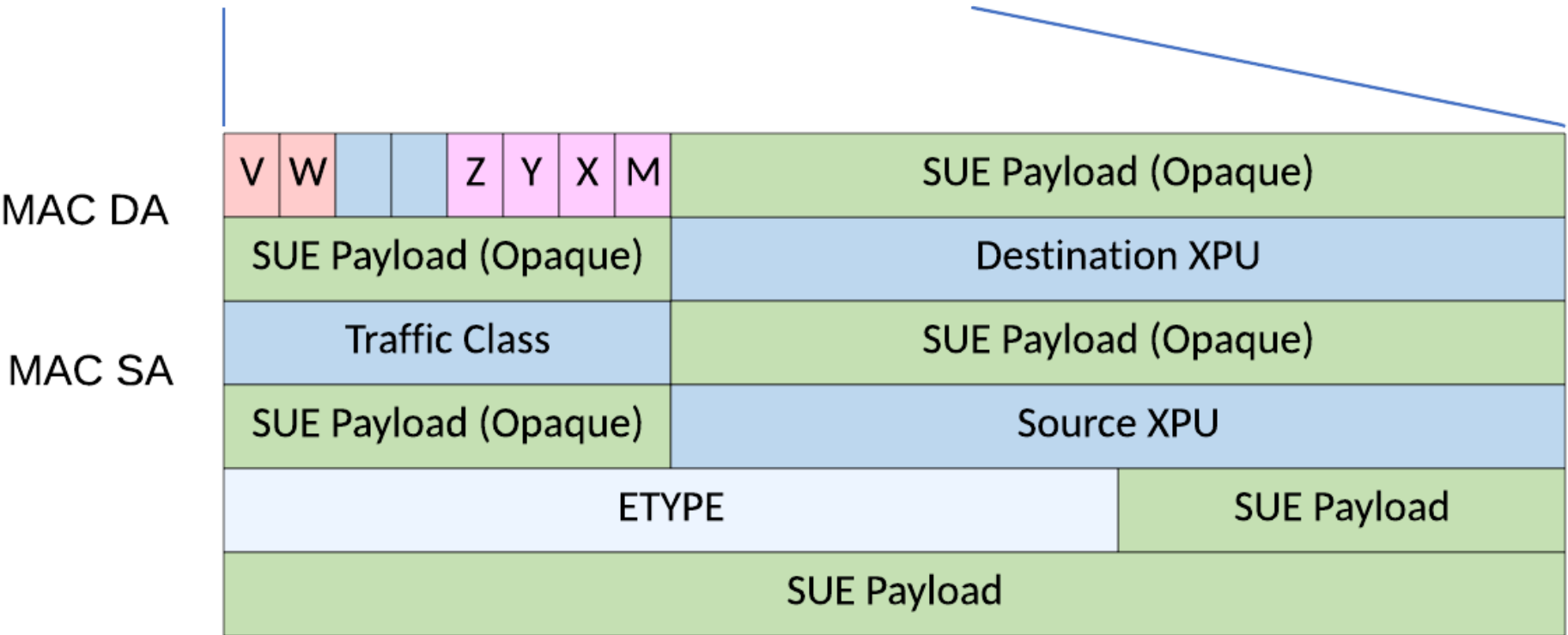
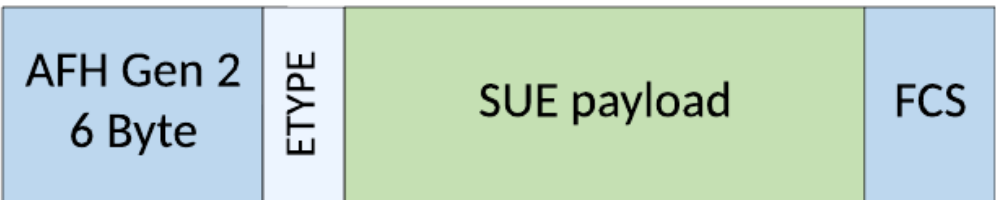
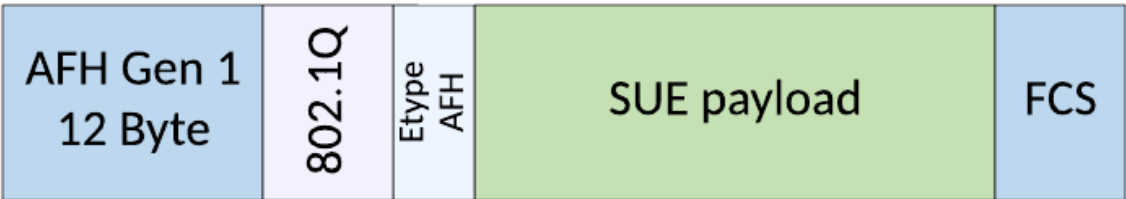
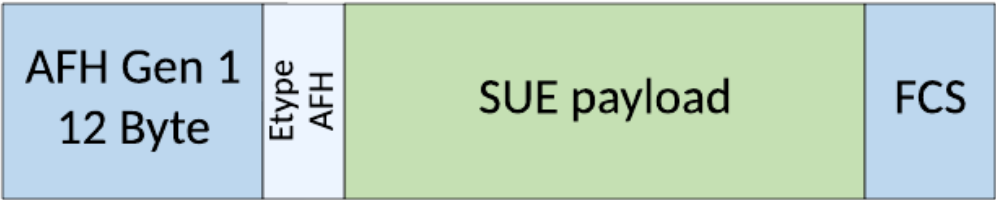
- ❖ Schedules a traffic class only if it has credits.

AI Fabric Header (AFH)

- **Motivations**
 - ❖ HPC and AI scale-up applications rely on memory load-store operations where payload size is small.
 - ❖ Ethernet + IP headers give flexibility, but it results in large headers.
 - ❖ Wire efficiency matters for throughput and header overhead needs to be minimized.
- **TU supports optimized “AI Fabric Header” (AFH)**
 - ❖ AFH is overlaid on the Ethernet MAC header and incorporates a minimal set of useful fields.
 - ❖ Address fields are typically 2B or 4B
 - ❖ Deployments can interpret the namespace as device IDs, local identifiers, or IP addresses.
 - ❖ Switch forwarding is agnostic and ignores remaining bytes which may contain other end-end information.
 - ❖ The approach retains full Ethernet MAC compatibility for easy adoption.



AFH for Scale Up Ethernet

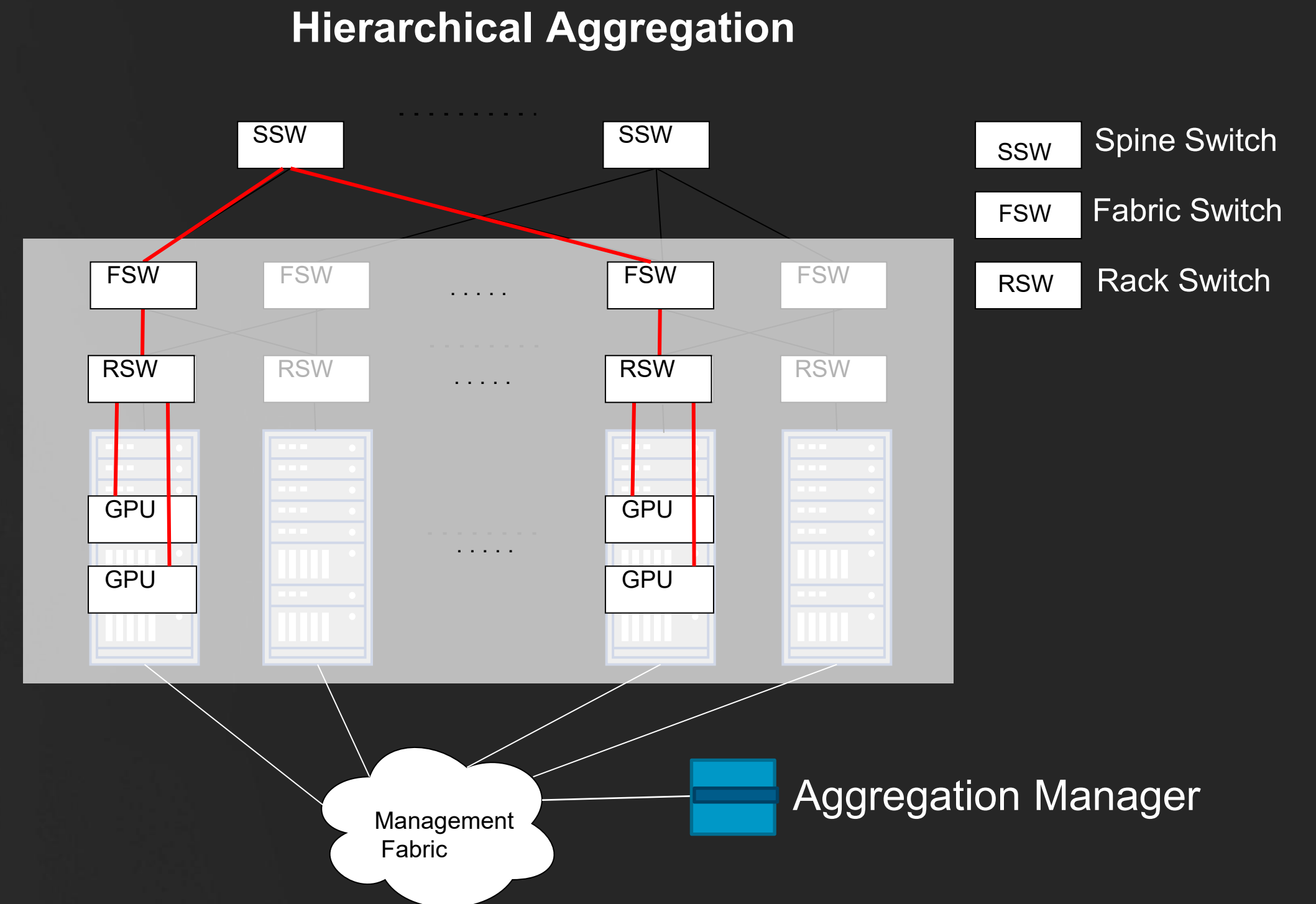


- ❖ Header identified by ZYXM bits as per IEEE SLAP spec.
- ❖ Two variations, one suitable for (2B address) and another for (4B address, entropy, hop count).
- ❖ Scale-up format uses 6B, rest available for end-end use.
- ❖ Switch identifies format and ignores other bytes.

See SUE Reference specification

TU – Support for Collective Operations (INC)

- In-Network Collectives (INC) offloads collective operation to TU
- High-level objectives
 - ❖ Work with current endpoints
 - ❖ Work with multiple chip architectures
 - ❖ Require minimal changes to current software models
- Example collectives supported:
 - ❖ Reduce
 - ❖ All Reduce
 - ❖ Broadcast
- Support Read or Write RDMA operations



Topology Aware Adaptive Routing

Topologies

Adaptive routing in TH-Ultra supports hierarchical, tree-like and mesh topologies

- ❖ Example, DragonFly+, CLOS, and 3D torus

AR Capabilities

Adaptive load aware selection of minimal, non-minimal and deterministic paths.

- ❖ Per VC metrics, forwarding policies and constraints to avoid deadlocks.
- ❖ Direction of packet arrival affects route FIB selection and subsequent path choices.
- ❖ Support for high-radix topologies.

Path Quality Metrics

Extensible path quality metrics

- ❖ Native support for per VC and aggregate port metrics.
- ❖ On-chip programmability to define new metrics based on on-chip measurements.
- ❖ VCs control which metrics to use. Example, can enable different metrics for lossy and lossless traffic.

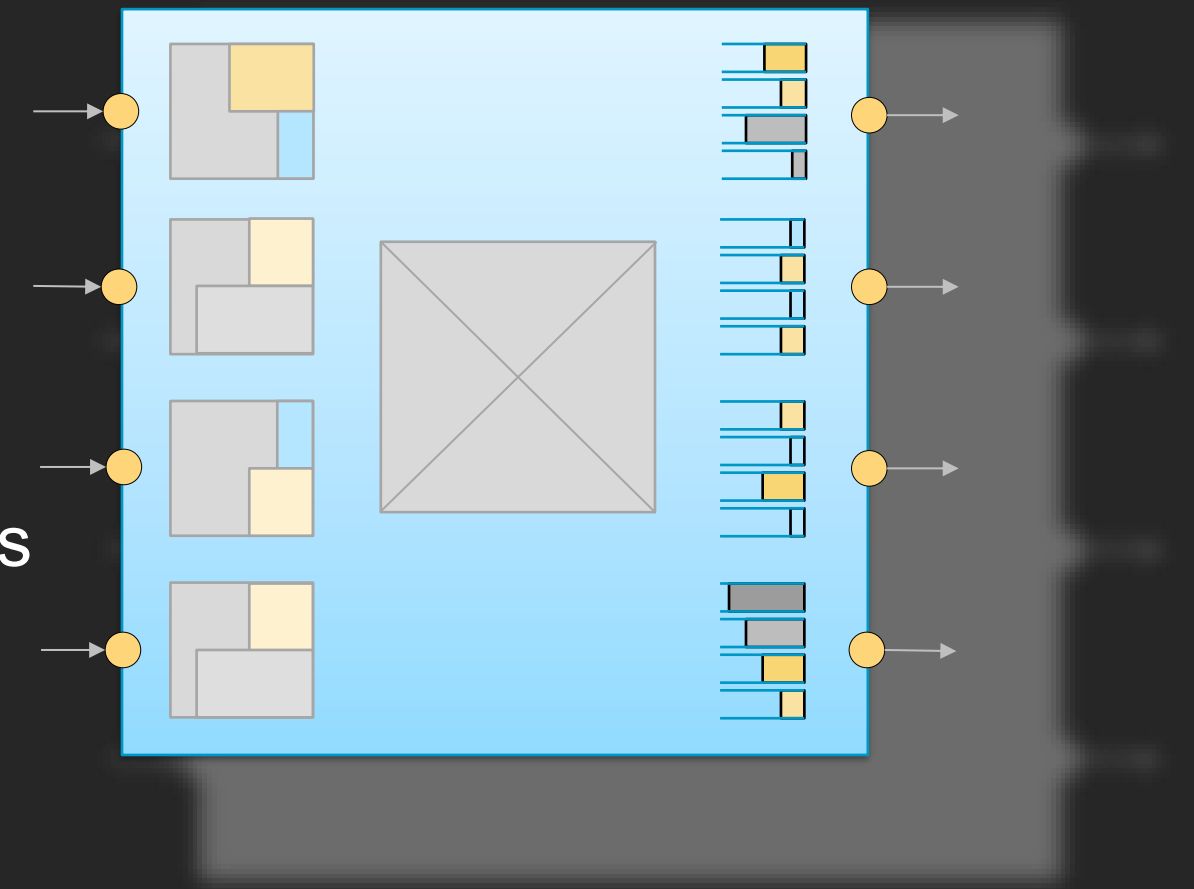
Forwarding Policies

Flexible forwarding policies

- ❖ Spray, Flowlet and Reactive Flowlet modes.
- ❖ Per-packet adaptive and deterministic resolution.
- ❖ Fast failover when links fail (< 500ns).
- ❖ Ability to maintain path affinity for a flow when load based policies select alternate paths.

Programmable Visibility

- Network behavior at micro-second level granularity
 - Traffic interference & performance issues not visible at large time scales
- Switch collects rich set of metrics
 - E.g., Queue lengths, packet delays, congestion stats, drops etc
 - Metrics are distributed in nature across the switch components
 - User based Python program to define metrics based on switch state such as queue lengths
- Visibility framework
 - Fast access to observations
 - Compute accelerator translates to actionable metrics
 - Metrics can be exported, acted locally and fast response time



Congestion Control

Proactive Telemetry

- ❖ In-band telemetry using CSIG and IFA

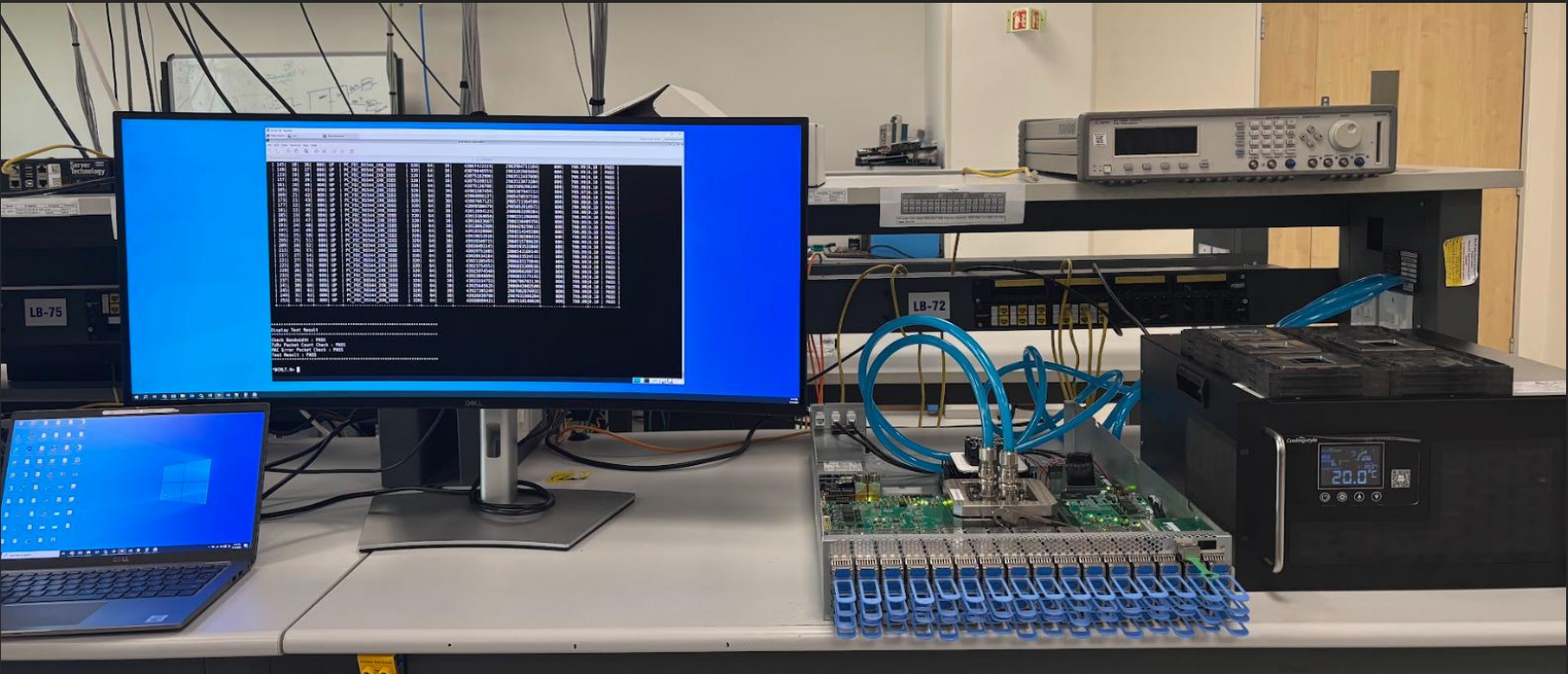
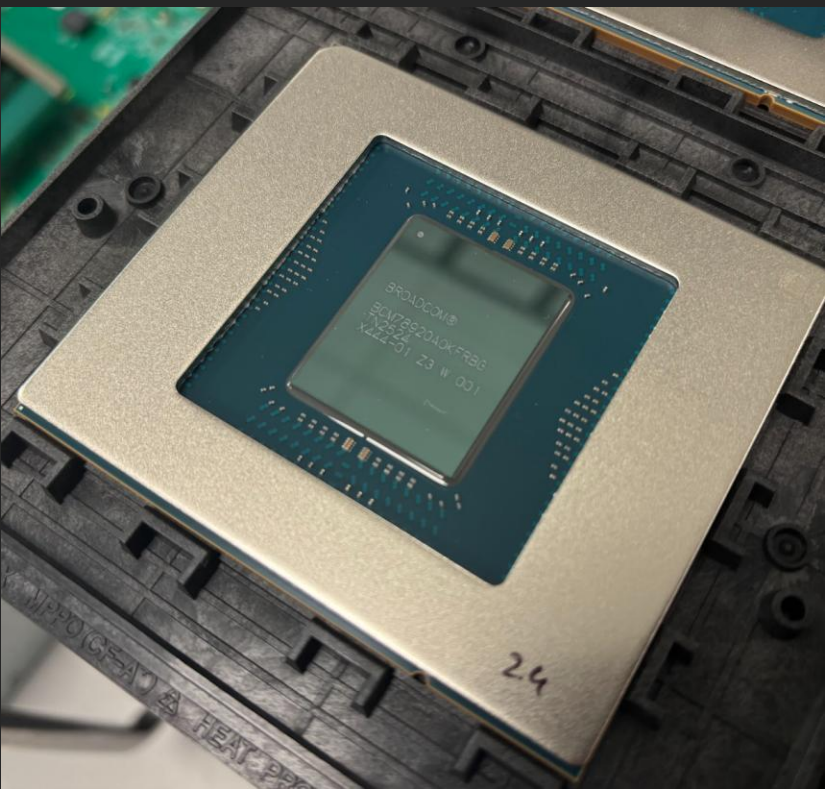
Reactive Feedback

- ❖ Packet trimming as defined by UEC 1.0.
- ❖ Fast congestion notification to endpoints, aka Back to Sender.
- ❖ ECN marking based on buffer occupancy and residence time.

Tomahawk Ultra Performance (1)

PktSz	Time (sec)	Total Pkts	Total Bytes	Exp.Rate (Gbps)	Obs.Rate (Gbps)
64	30	42876153178	2915578416104	800	799.99
64	30	42876922472	2915630728096	800	799.99
64	30	42879084392	2915777738656	800	799.99
64	30	42879184021	2915784513428	800	799.99
64	30	43279114690	2942979798920	800	799.97
64	30	43284217326	2943326778168	800	799.99
64	30	43289563949	2943690348532	800	799.99
64	30	43290198852	2943733521936	800	799.99
64	30	43290689501	2943766886068	800	799.99
64	30	43293097582	2943930635576	800	799.99
64	30	43294384425	2944018140900	800	799.99
64	30	43296920187	2944190572716	800	799.99
64	30	43297910995	2944257947660	800	799.99
64	30	43300131355	2944408932140	800	799.99
64	30	43301764787	2944520005516	800	799.99
64	30	43302575458	2944575131144	800	799.99
64	30	43303395516	2944630895088	800	799.99
64	30	43304192682	2944685102376	800	799.99
64	30	43304135586	2944681219848	800	799.99
64	30	43303536082	2944640453576	800	799.99
64	30	43304256696	2944689455328	800	799.99
64	30	43303355608	2944628181344	800	799.99
64	30	43304211274	2944686366632	800	799.99
64	30	43304223620	2944687206160	800	799.99
64	30	43306237255	2944824133340	800	799.99
64	30	43308006873	2944944467364	800	799.99
64	30	43305978615	2944806545820	800	799.99
64	30	43305350239	2944763816252	800	799.99
64	30	43307250806	2944893054808	800	799.99
64	30	43308722178	2944993108104	800	799.99
64	30	43308174785	2944955885380	800	799.99
64	30	43308956075	2945009013100	800	799.99
64	30	43309719314	2945060913352	800	799.99
64	30			800	799.99

Line rate @64B
on all ports



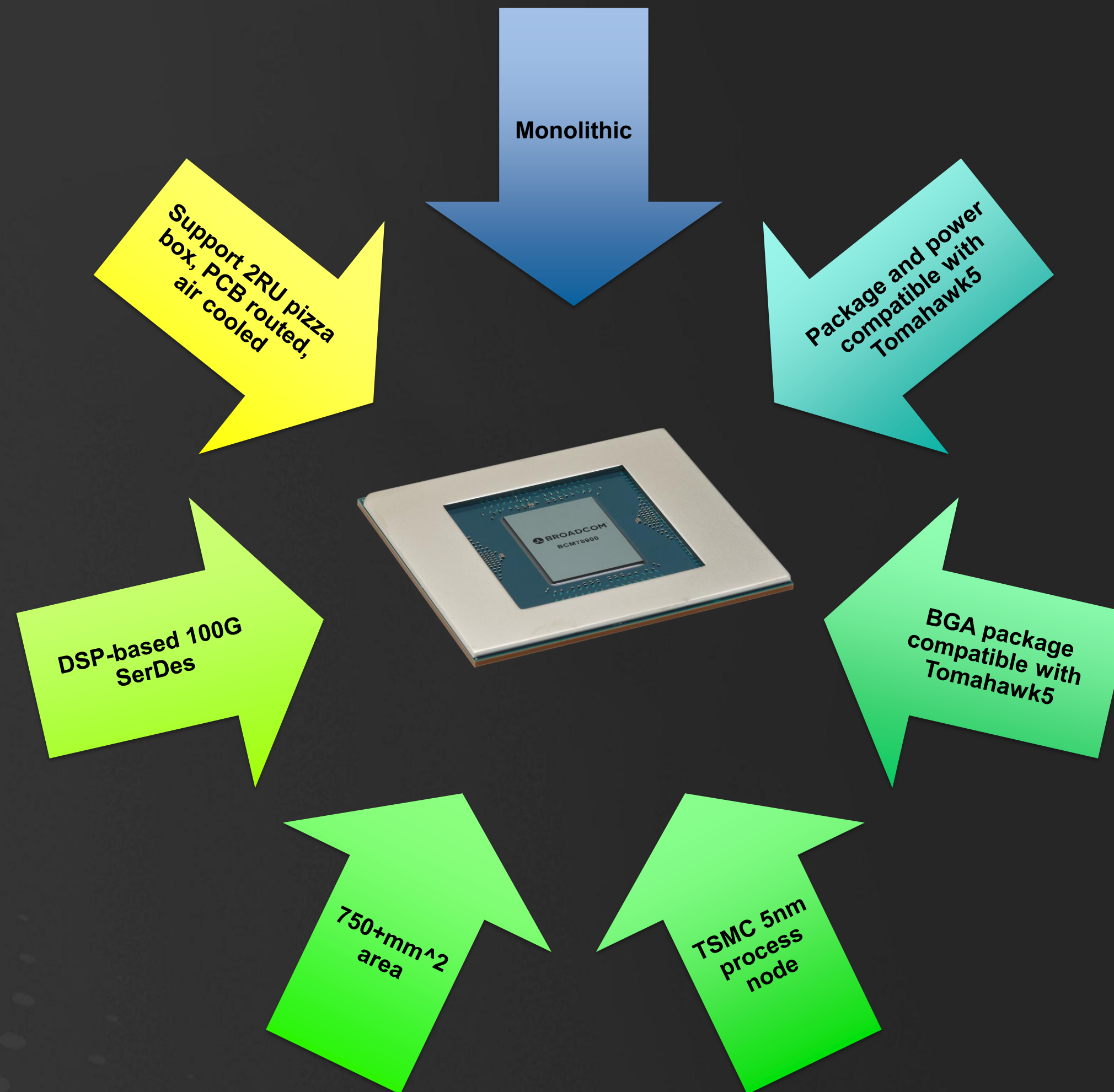
Tomahawk Ultra Performance (2)

Tsc No.	Port No.	Latency (ns)							
		Min	Max	Avg	Max	p90	p95	p99	p999
					Min				
TSC0	1	239	241	240.46	2	241	241	241	241
TSC1	5	241	254	245.24	13	249	250	254	254
TSC2	9	237	239	238.08	2	239	239	239	239
TSC3	13	237	246	238.72	9	239	240	245	246
TSC4	17	241	243	242.05	2	243	243	243	243
TSC5	21	238	240	238.83	2	239	240	240	240
TSC6	25	232	235	233.39	3	234	234	235	235
TSC7	29	233	235	233.62	2	234	234	235	235
TSC8	33	235	238	236.92	3	238	238	238	238
TSC9	37	235	237	236.26	2	237	237	237	237
TSC10	41	235	238	236.61	3	237	237	238	238
TSC11	45	232	235	233.70	3	234	235	235	235
TSC12	49	232	235	233.49	3	234	234	235	235
TSC13	53	230	232	231.09	2	232	232	232	232
TSC14	57	227	229	228.15	2	229	229	229	229
TSC15	61	227	231	229.05	4	230	230	230	231
TSC16	65	226	229	227.78	3	229	229	229	229
TSC17	69	227	230	228.79	3	229	230	230	230
TSC18	73	230	232	231.21	2	232	232	232	232
TSC19	77	232	234	233.04	2	234	234	234	234
TSC20	81	233	236	234.33	3	235	235	236	236
TSC21	85	236	238	236.96	2	238	238	238	238
TSC22	89	235	238	236.35	3	237	237	237	238
TSC23	93	236	238	237.38	2	238	238	238	238
TSC24	97	234	236	235.25	2	236	236	236	236
TSC25	101	233	236	234.65	3	235	236	236	236
TSC26	105	239	241	239.98	2	241	241	241	241
TSC27	109	241	244	242.52	3	243	243	244	244
TSC28	113	240	243	241.40	3	242	242	243	243
TSC29	117	238	240	239.04	2	240	240	240	240
TSC30	121	240	242	240.66	2	241	242	242	242
TSC31	125	239	242	240.20	3	241	241	241	242
TSC32	129	239	241	239.73	2	240	241	241	241
TSC33	133	241	243	241.82	2	243	243	243	243
TSC34	137	237	240	238.33	3	239	239	240	240
TSC35	141	239	242	240.46	3	241	241	242	242
TSC36	145	239	241	240.30	2	241	241	241	241
TSC37	149	239	241	240.14	2	241	241	241	241
TSC38	153	233	236	234.55	3	235	235	236	236
TSC39	157	234	236	235.18	2	236	236	236	236
TSC40	161	238	241	239.08	3	240	240	240	241
TSC41	165	234	237	235.94	3	237	237	237	237
TSC42	169	233	236	234.65	3	235	235	236	236
TSC43	173	231	234	232.10	3	233	233	233	234
TSC44	177	233	236	234.87	3	236	236	236	236
TSC45	181	229	231	230.07	2	231	231	231	231
TSC46	185	226	229	227.31	3	228	228	229	229
TSC47	189	227	229	227.80	2	228	229	229	229
TSC48	193	228	230	228.91	2	229	230	230	230
TSC49	197	228	230	229.24	2	230	230	230	230
TSC50	201	229	231	229.62	2	230	231	231	231
TSC51	205	233	235	234.15	2	235	235	235	235
TSC52	209	231	233	231.87	2	233	233	233	233
TSC53	213	233	235	233.84	2	235	235	235	235
TSC54	217	234	237	235.68	3	236	237	237	237
TSC55	221	237	240	239.03	3	240	240	240	240
TSC56	225	233	235	234.35	2	235	235	235	235
TSC57	229	232	235	233.23	3	234	234	234	235
TSC58	233	237	239	238.12	2	239	239	239	239
TSC59	237	242	245	243.02	3	244	244	244	245
TSC60	241	239	242	240.46	3	241	241	242	242
TSC61	245	238	240	239.16	2	240	240	240	240
TSC62	249	239	241	239.98	2	241	241	241	241
TSC63	253	239	242	240.43	3	241	241	241	242

Tsc No.	Port No.	S	Latency (ns)							
			Min	Max	Avg	Max	p90	p95	p99	p999
						Min				
TSC0	1		239	241	240.46	2	241	241	241	241
TSC1	5		241	254	245.24	13	249	250	254	254
TSC2	9		237	239	238.08	2	239	239	239	239
TSC3	13		237	246	238.72	9	239	240	245	246
TSC4	17		241	243	242.05	2	243	243	243	243
TSC5	21		238	240	238.83	2	239	240	240	240
TSC6	25		232	235	233.39	3	234	234	235	235
TSC7	29		233	235	233.62	2	234	234	235	235
TSC8	33		235	238	236.92	3	238	238	238	238
TSC9	37		235	237	236.26	2	237	237	237	237
TSC10	41		235	238	236.61	3	237	237	238	238
TSC11	45		232	235	233.70	3	234	235	235	235
TSC12	49		232	235	233.49	3	234	234	235	235
TSC13	53		230	232	231.09	2	232	232	232	232
TSC14	57		227	229	228.15	2	229	229	229	229
TSC15	61		227	231	229.05	4	230	230	230	231
TSC16	65		226	229	227.78	3	229	229	229	229
TSC17	69		227	230	228.79	3	229	230	230	230
TSC18	73		230	232	231.21	2	232	232	232	232
TSC19	77		232	234	233.04	2	234	234	234	234
TSC20	81		233	236	234.33	3	235	235	236	236
TSC21	85		236	238	236.96	2	238	238	238	238

Latency < 250 ns

Implementation Highlights



Broadcom Switch Coverage of HPC and AI

HPC



Tomahawk Ultra

51.2 Tbps

AI Scale-Up



Tomahawk 6

102.4 Tbps

AI Scale-Out



Jericho4

51.2 Tbps

Region Scale-Out

Summary



TOMAHAWK-ULTRA BRINGS
ETHERNET TO SCALEUP AND
HPC MARKETS



77B PPS RATE



< 250NS LATENCY



LOSSLESS OPERATION VIA
CBFC AND PFC



IMPROVED LINK RELIABILITY
VIA LLR



HIGH BANDWIDTH IN-
NETWORK-COLLECTIVES



PACKAGE AND POWER
COMPATIBLE WITH
TOMAHAWK5 FOR FAST TTM

Tomahawk Ultra – Ultra Low Latency, High Bandwidth Ethernet Switch for HPC & AI/ML applications

Thank you

OPEN // SCALABLE // POWER EFFICIENT