CM Mechanisms: NIC perspective

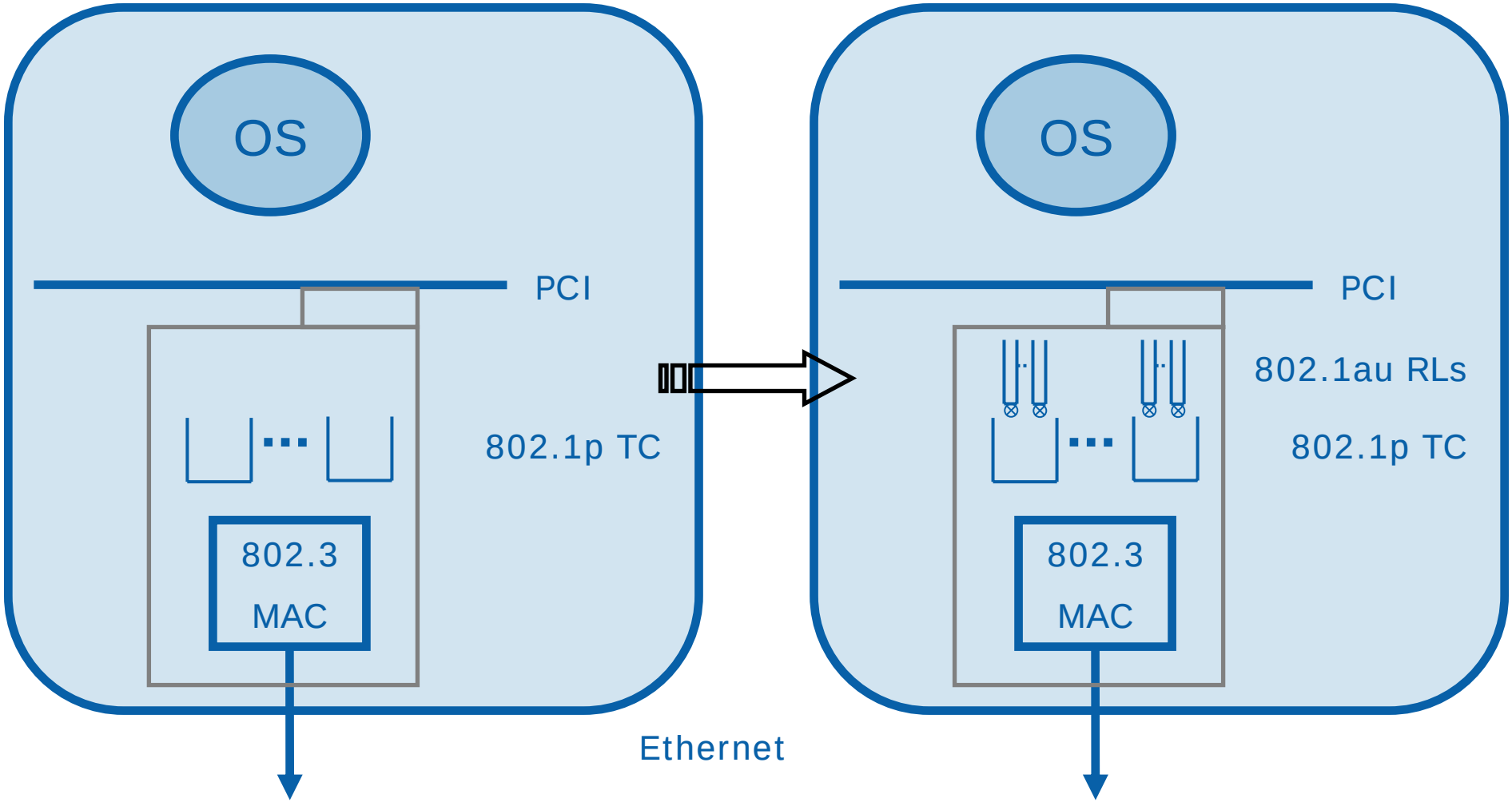
Manoj Wadekar (Intel)

Pat Thaler (Broadcom)

Agenda

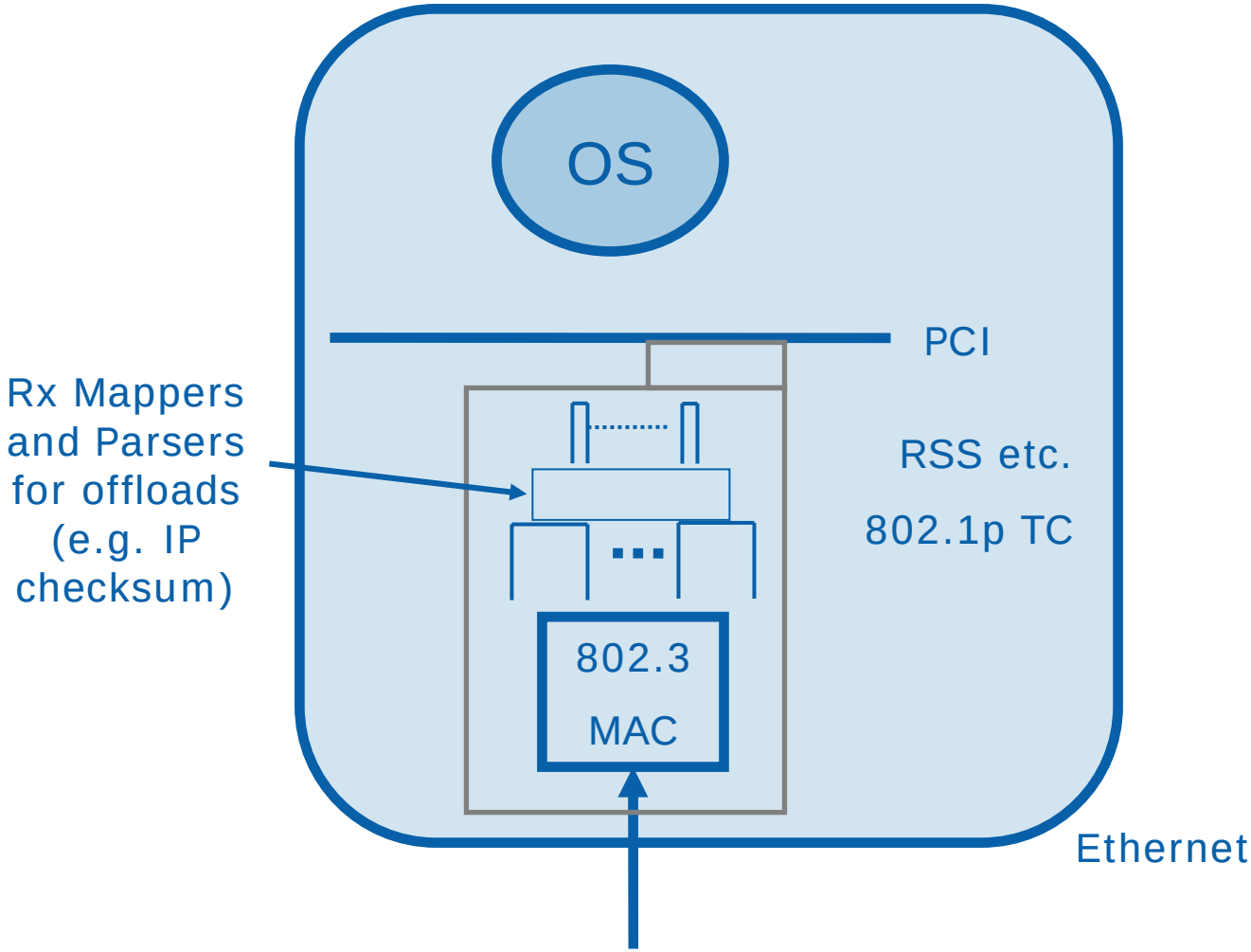
- NIC design impact of CM mechanism
- Comparison of proposed solutions
- Management challenges
- Summary Recommendations

NIC CN impact: Transmit



Allow practical cost-performance tradeoff

NIC CN impact: Receive



Tags can add complexity in Receive side processing

NIC: CM protocol impact

- CM packet consumption:
  - Need to respond to CM messages received by reducing RL rate on Tx-side
  - Needs real time handling
  - Less signaling is better
- Reflection in NICs:
  - NICs (typically) do not forward packets from Rx to Tx
  - Reflection of tags requires forwarding fragment of Rx packet to Tx
  - If reflection is required, packet to packet is preferred over tag to packet
  - Protocol that needs NICs to perform reflection will add burden
- Per RL state in Transmit:
  - E.g. rate recovery, drift require timer/packet based state machine per RL
  - Simpler the better
  - Should balance performance-complexity trade-off

2-point CM proposals

- Poll during July Plenary meeting provided TF preference for 2-point CM proposals
  - 3-point may provide some optimization but adds too much complexity
  - 3-point has issues in presence of multi-path and flow coalescing
- Currently BCN and QCN are 2-point CM proposals
- Both mechanisms have many similarities
  - CP provides congestion information to RPs via CN messages
    - Based on Q and Q-derivative
  - RP's algorithmically adapt their transmission rate to match capacity
    - Distributed congestion management vs. rate allocation by CP
  - Response to congestion event is similar for both mechanisms
  - Similar configuration parameters for both mechanisms
- Proposals differ primarily in their rate recovery mechanisms

Comparison of QCN and BCN proposals

Feature	BCN	QCN	RP Impact
Fb computation point	At Reaction Point (RP)	At Congestion Point (CP)	QCN Simplifies RP
Feedback parameters sent	Qoff, Qdelta (16b, 16b)	Quantized Fb (6b)	QCN Simplifies RP
Sampling Probability	Fixed w/ Jitter, over-sampling	Fb-based sampling (1% to 10%)	QCN Increases signaling during congestion
Positive feedback	Yes, Additive increase proportional to Fb	No	QCN Lowers signaling
RP-CP association in RL-Tag	Yes, as required for positive feedback	No	QCN simplifies RP
Trigger to self-recovery	Time-based drift	Packet-based & Time-based	QCN adds complexity to RP
Emergency feedback	BCN0, BCN(Max)	No (Fb-based sampling, Fb w/ better range)	BCN(0), BCN(Max) Valuable if it lowers signaling

QCN provides complexity-performance trade-off

Comparison of QCN and BCN parameters

BCN		QCN	
Qeq	CP	Qeq	CP
W	RP	W	CP
Gd	RP	Gd	RP
Rmin	RP	Rmin	RP
driftFactor	RP	driftFactor	RP
driftTimer	RP	driftTimer	RP
Gi	RP	toThreshold	RP
samplingInterval	CP	baseProbability	CP
Qsc	CP	extraFastRecovery	RP
Qmc	CP	fastRecoveryThreshold	RP
Qsat	CP	hyperactiveIncrease	RP
Rd	RP	minDecFactor	RP
BCN0	CP	EfrMax	RP
		A (additive increase)	RP
6 parameters at CP and 7 at RP		**3 parameters at CP and 11 at RP**	

Simplified default parameter template is necessary

Simplified Parameter Template

- Identify constants and variables from parameter list
- Reduce number of variables: primary/derived
  - E.g. Qsc, Qmc derived from Qeq, Rmin derived from C etc.
- Reduce number of (experimental) options
- Define default values for majority of variables:
  - That work for large control loop delay up to 500 uS (with certain acceptance criteria)
  - That work for large number of RPs : say, 1000
  - That allow reasonable and practical implementations
- Minimize deployment-specific variables:
  - E.g. Is W for large control loop delay is different than small CLD?
- Number of End Stations is (typically) larger than number of switches in Data Center
  - Avoid End Station tuning as much as possible

Summary Recommendations (from NIC perspective)

- QCN (2-point) provides good-enough solution
  - Leverages BCN fundamentals
  - Improves implementation by reducing signaling, avoiding tagging
- Simplicity of solutions should be maintained
  - Adopt framework and reduce low ROI options
- More attention should be provided now to reduce parameters
  - Validation/Management can be a challenge to CN deployment
  - Discovery protocol may allow switches to distribute configuration parameters to End Stations