Feedback Request:
DCB for HPC and Business Analytics

Clusters, MPI and Arbitrage Trading

Mitch Gusat
IBM Research, ZRL GmbH

Denver 2008

DCB From HPC & Business Analytics

- ETS: good
  - Open: traffic types, grouping and scheduling disciplines are TBD
- PFC: mixed
  - must have for HPC; PFC ~ VC's
  - TBD for Analytics ; PFC ~ VL's
  - Obs.: traffic separation = good
- QCN: mixed
  - no for HPC
  - no/TBD for BA
  - Option: separation of RP functionality (unwanted) from CP feedback (wanted)

Aggregated Req'ts: No Drop. What else?

- Lossless is a must => done!

Not done yet

1. Latency is primary metric in HPC and BA
   - interplay of scheduling, routing, flow control and stack

2. Network status info (aka feedback beyond congestion)
   1. timely: L2 is faster than Netflow and IPFix
   2. accurate: Qsize, Qeq, Qoff, Qdelta... already known by CP (but not by apps)

3. Bubble the L2 feedback up the L3/4/... stack
   1. see FlowID, RPID, KarmaID

Pervasive Feedback: Fb_Rq Option

- Performance monitoring, network profiling, runtime load balancing, adaptive routing… is there a single solution?
  - Feedback Request: **Fb_Rq**
- What is Fb_Rq?
- On demand status info, irrespective of congestion

Steps

1. SRC injects special (TBD) Fb_Rq pkt w/ L2 flag and SeqID
2. CP receives Fb_Rq
   1. set Psample=1 (or disregards if busy)
   2. dumps available L2 data: prio, Qoff, Qdelta, etc…
   3. sends Fb_Rp back to originating SRC
   4. [forwards Rq_RQ if the DST != local CP ➔ path profiling w/ one ping]

Thank you.