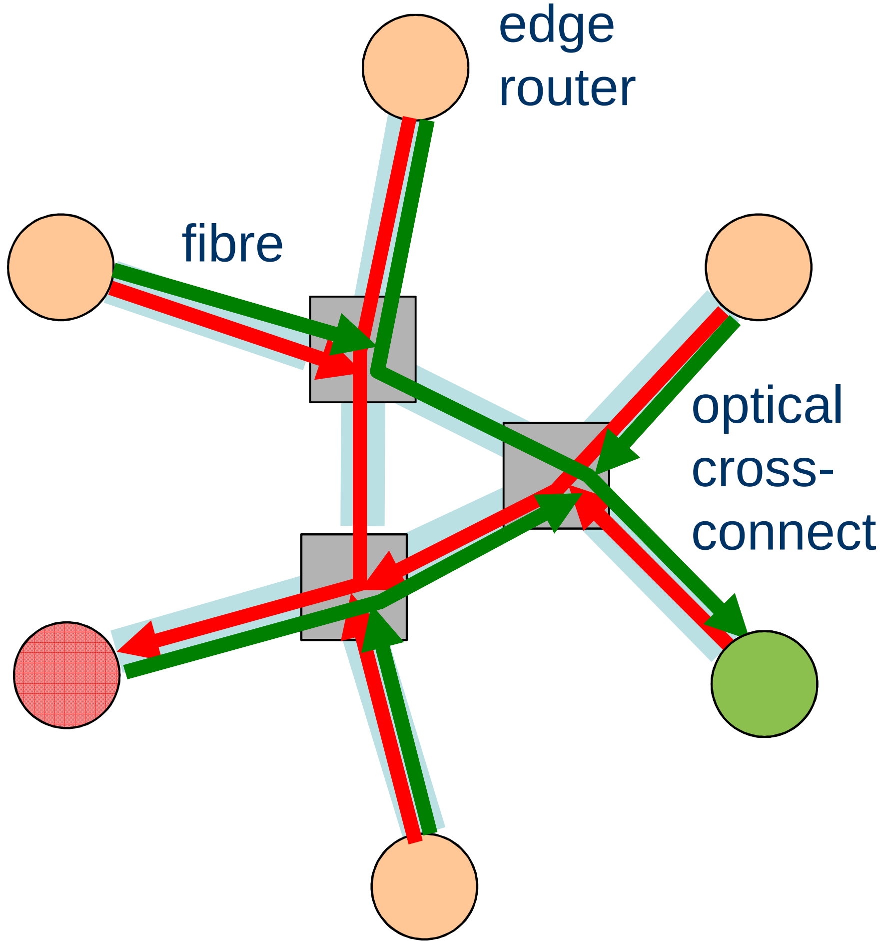
A flow-aware MAC protocol for a passive optical MAN

Philippe Robert and Jim Roberts
INRIA

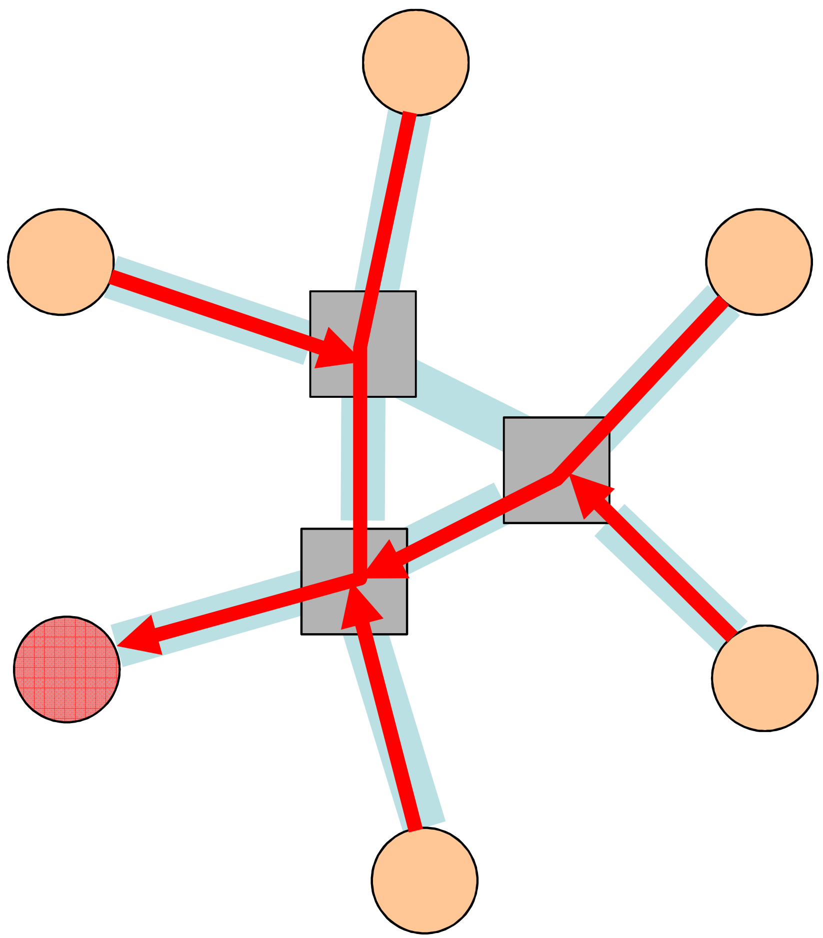
ITC 23
September 2011

A flow-aware MAC protocol for
a passive optical MAN



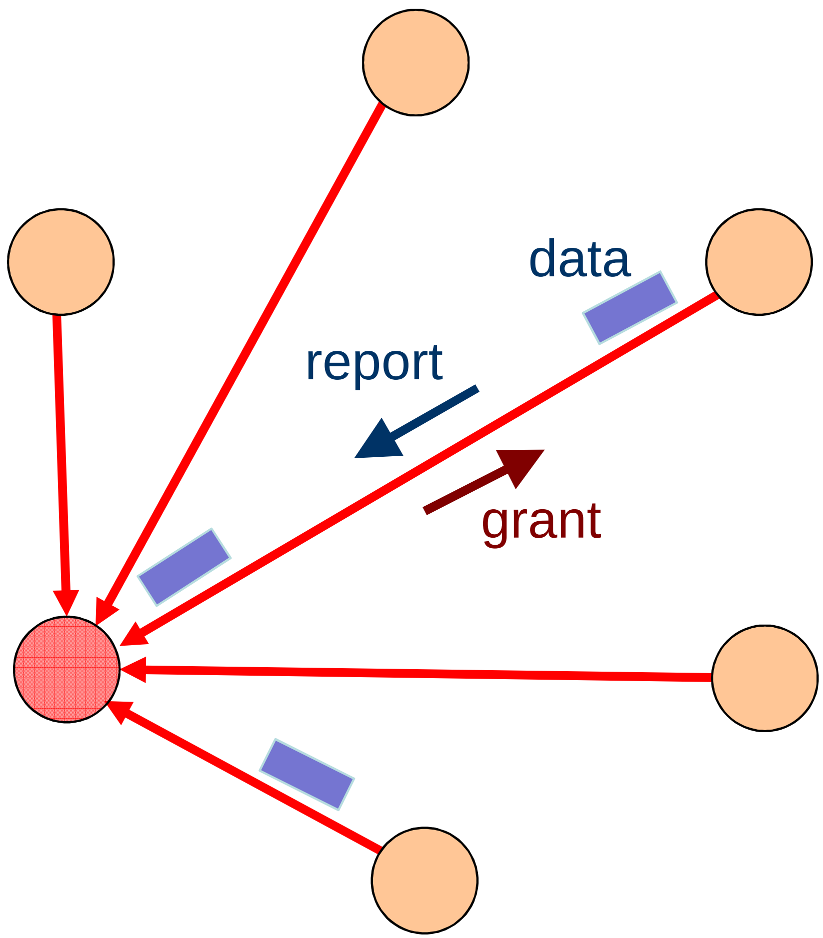
- using TWIN...
  - Widjaja et al. 2003
- to share lightpaths...
  - wavelength selective optical cross-connects, tunable transmitters and burst mode receivers
- in a metropolitan area network (MAN)
  - aggregated traffic, short distances

A flow-aware MAC protocol for a passive optical MAN



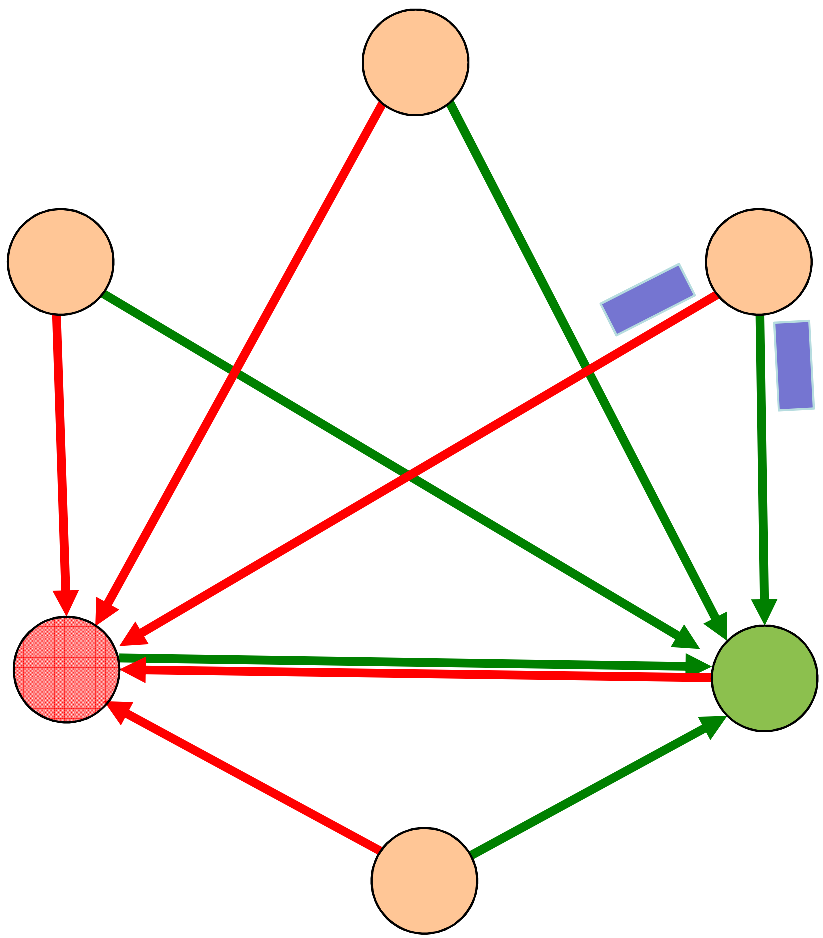
- burst timing to avoid collisions
  - at destinations and at cross-connects

A flow-aware MAC protocol for a passive optical MAN



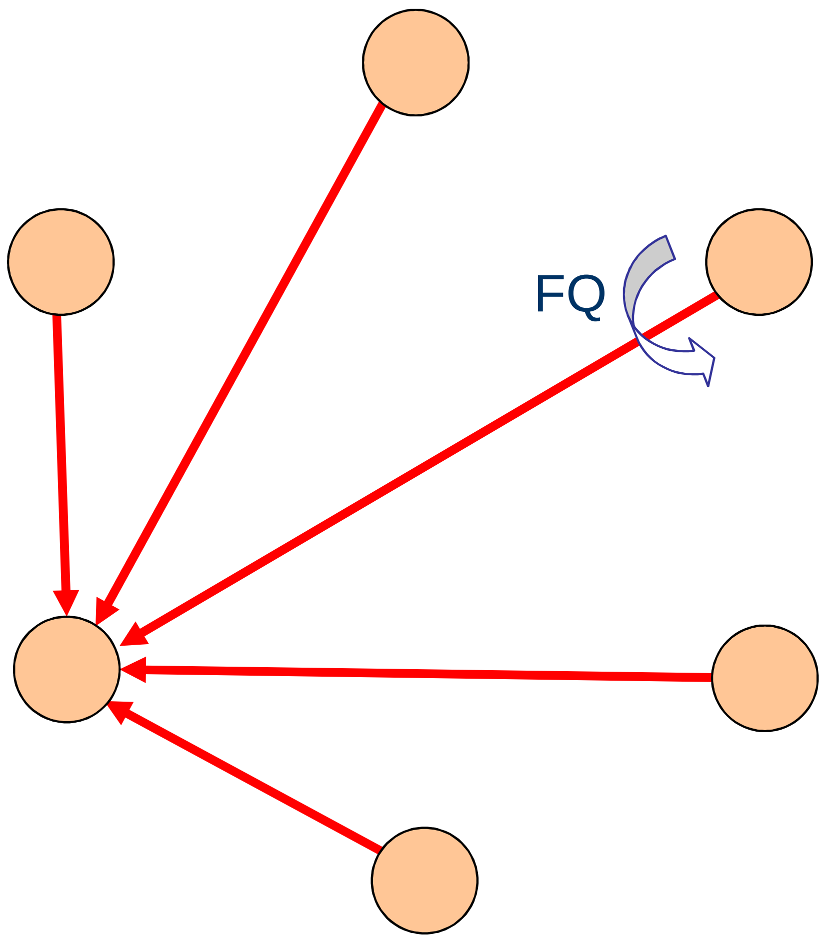
- burst timing to avoid collisions
  - at destinations and at cross-connects
- sources send reports
  - current queue contents
- destination gives grants
  - that do not collide

A flow-aware MAC protocol for a passive optical MAN



- burst timing to avoid collisions
  - at destinations and at cross-connects
- sources send reports
  - current queue content
- destination allocates grants
  - that do not collide
- but grants suffer from transmitter blocking

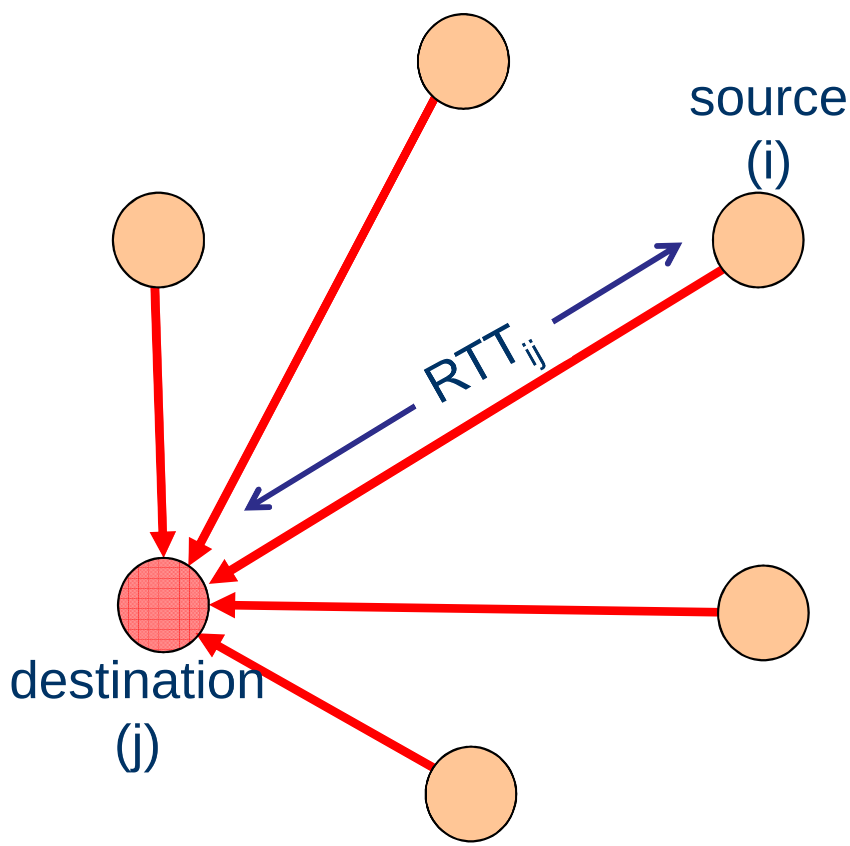
A flow-aware MAC protocol for a passive optical MAN



- grants size proportional to number of active flows
- per-flow fair queueing and overload control
  - scalable and feasible
- for implicit service differentiation
- and a transport agnostic network

Timing grants to avoid collision

- synchronization and ranging as in EPON
  - synchronize source i and destination j clocks
  - destination measures round trip time RTT_{ij}



Timing grants to avoid collision

- synchronization and ranging as in EPON
  - synchronize source i and destination j clocks
  - destination measures round trip time RTT_{ij}
- destination j computes n^{th} grant recursively

$$g(n) = g(n-1) + d(n-1) + \Delta_R$$
$$s(n) = g(n) + \Delta_O - RTT_{ij}$$

- $g(n)$ is when n^{th} grant is computed,
- $s(n)$ is start time on source i clock,
- $d(n-1)$ is $n-1^{th}$ grant duration,

  Δ_R is guard time + report time,

  Δ_O is *large enough* offset (max $\{RTT_{ij}\} + \tau$)

Timing grants to avoid collision

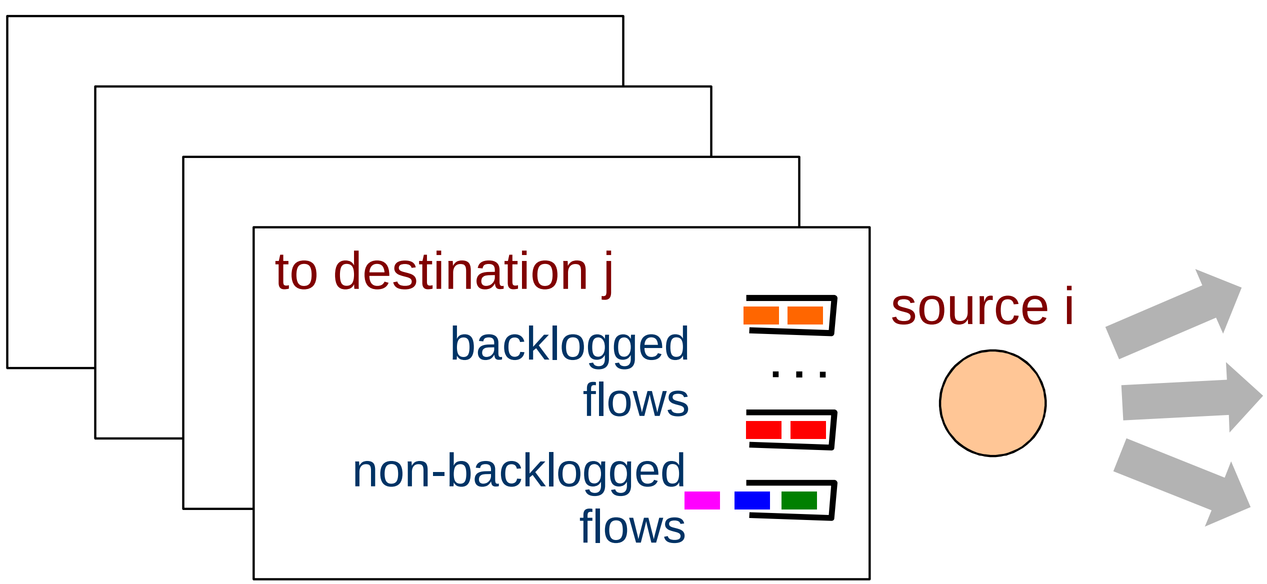
- synchronization and ranging as in EPON
  - synchronize source i and destination j clocks
  - destination measures round trip time RTT_{ij}
- destination j computes n^{th} grant recursively

$$g(n) = g(n-1) + d(n-1) + \Delta_R$$
$$s(n) = g(n) + \Delta_O - RTT_{ij}$$

- provably efficient and feasible
  - i.e., fully uses capacity, avoids collisions, grants arrive in time
- reports signalled in-band, grants signalled out-of-band
- choice of service order and grant size $d(n)$ is open

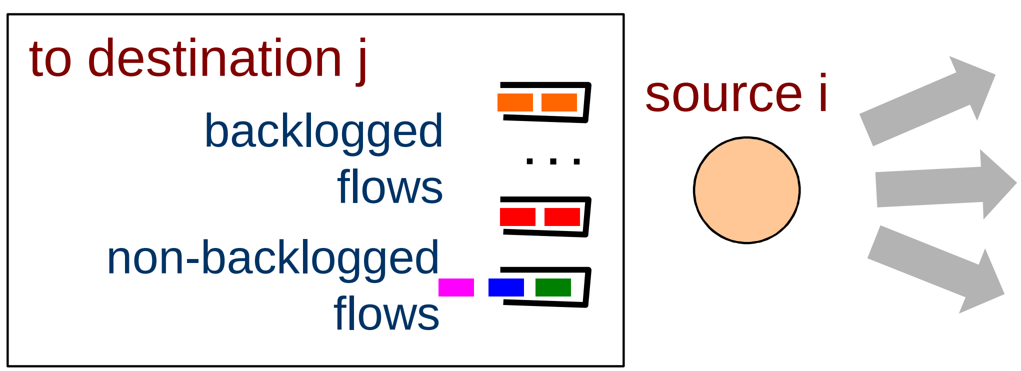
Flow-aware reports and grants

- sources implement "priority deficit round robin"
  - fair sharing between backlogged flows
  - priority to packets of non-backlogged flows
  - cf. Kortebi et al., 2005



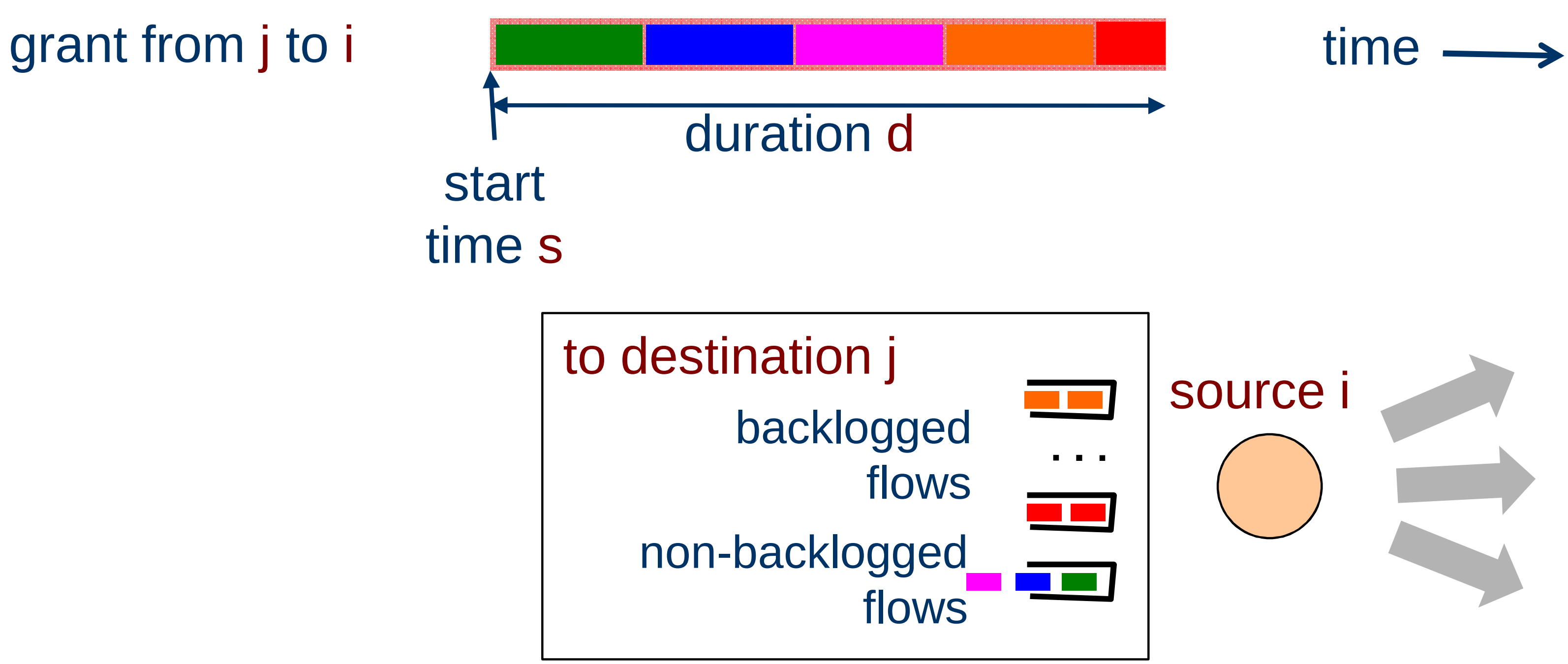
Flow-aware reports and grants

- sources implement "priority deficit round robin"
  - fair sharing between backlogged flows
  - priority to packets of non-backlogged flows
  - cf. Kortebi et al., 2005
- report (i,j) ⇒ number of backlogged flows, size of non-backlogged flow queue
- grant (i,j) ⇒ 1 "quantum" for each backlogged flow + latest reported priority queue size



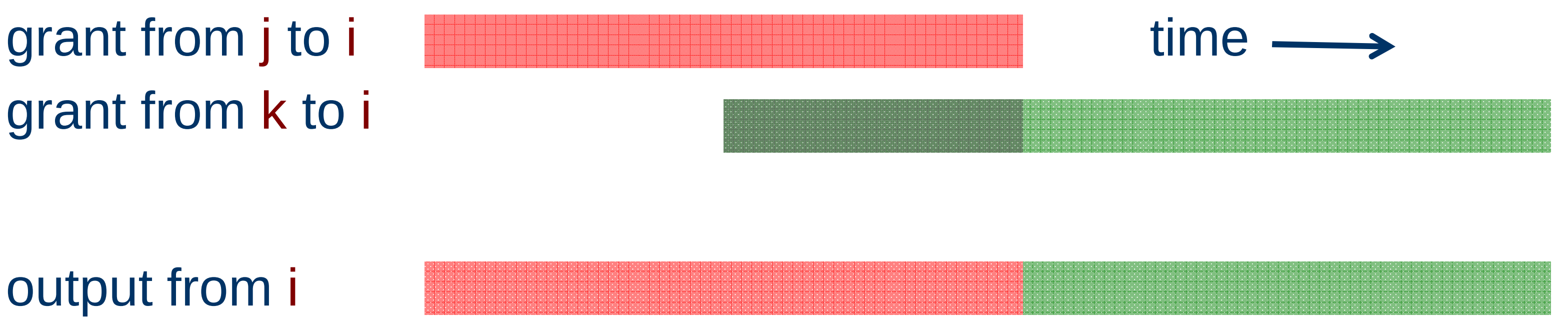
Filling grants

- queue contents change between report epoch and grant start time
  - include all waiting packets in priority queue, fill up with quanta from backlogged flows



Filling grants

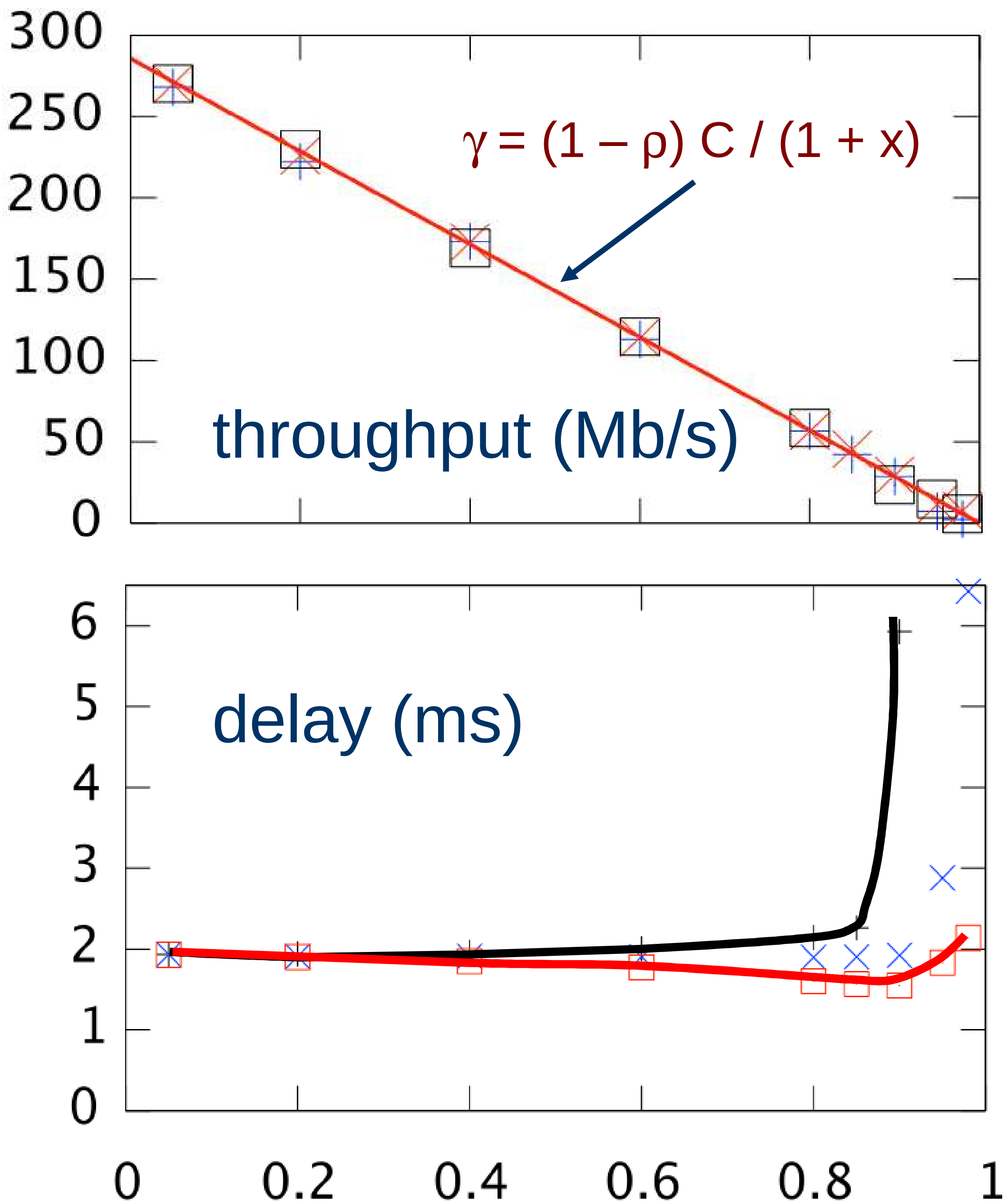
- queue contents change between report epoch and grant start time
  - include all waiting packets in priority queue, fill up with quanta from backlogged flows
- when transmitter blocking occurs, use grants in start time order without pre-emption
  - account for lost grant time in next report



Performance of one lightpath

- traffic capacity is maximal, ie, number of backlogged flows is stable if and only if
  - demand (= arrival rate x size) < wavelength capacity (ie, $\rho < 1$)
  - proof by Lyapunov function
- approximation by processor sharing model
  - assume all flows are backlogged
  - consider limit quantum $\rightarrow 0$, overhead = nodes$*\Delta_R$ = x$*$quantum
  - a PS queue with a permanent rate x customer
- expected flow throughput, $\gamma = (1 - \rho)\, C / (1 + x)$
- a reduced load approximation to account for non-backlogged flows ($\Rightarrow$ same γ)

Simulation results for 10 x 10 MAN (1 Gb/s)



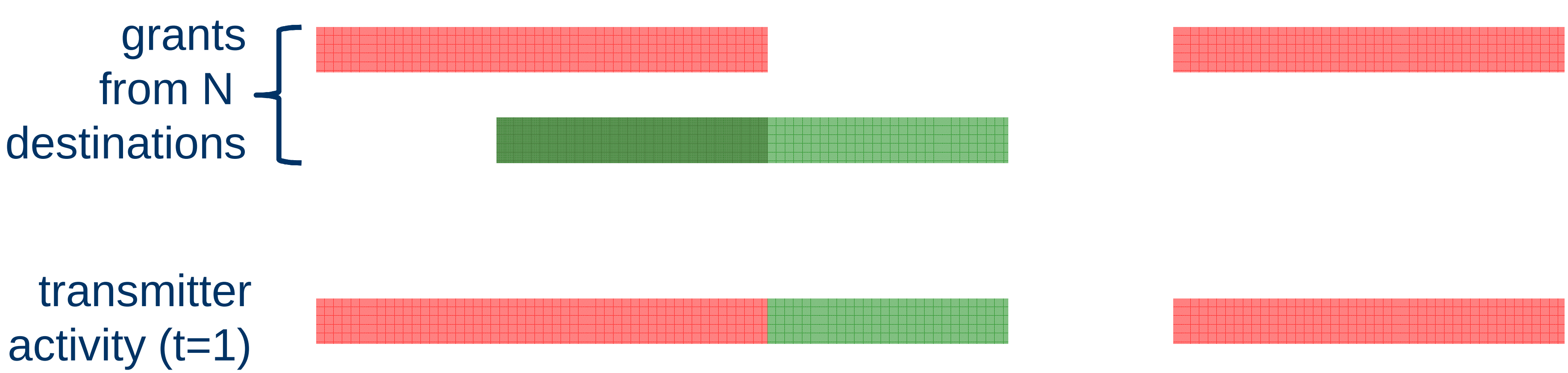
- traffic mix
  - 20% backlogged +
  - 60% backlogged ×
  - 100% backlogged □
- confirms maximal capacity
- significant overhead at low load
- traffic mix
  - 0% backlogged +
  - 20% backlogged ×
  - 60% backlogged □
- negligible delay until saturation

Multiple trees: accounting for transmitter blocking

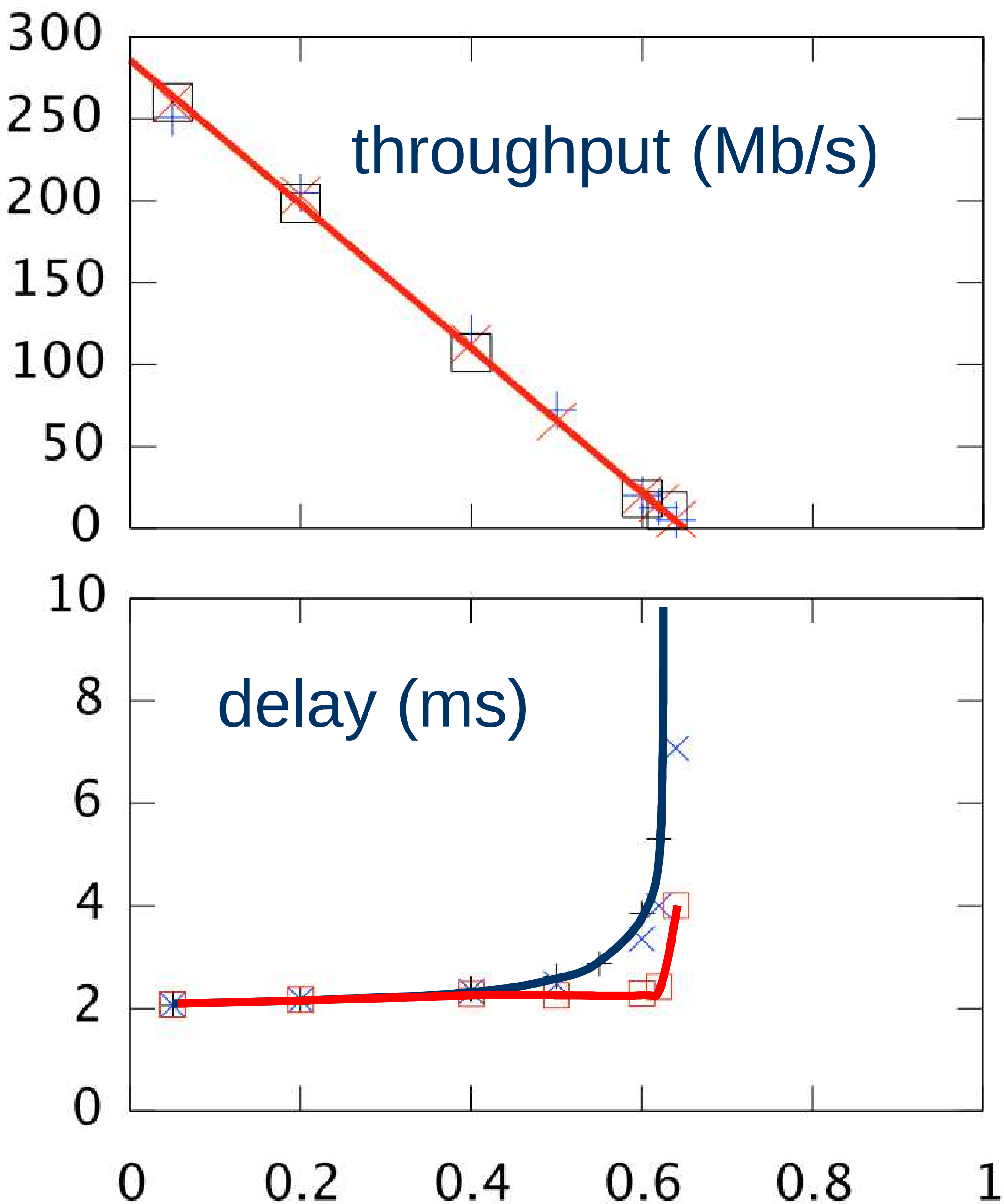
- proportion of lost capacity with t transmitters, $B_t(\rho)$, is given by the Engset formula (assuming independence)
- deduce load ρ^* at which transmitters fully busy
- $\Rightarrow$ traffic capacity is reduced by $(1-B_t(\rho^*))$

$$\gamma H (1 - \rho/(1-B_t(\rho^*)) C / (1 + x)$$

- for ≥ 10 node network, $B_1(\rho^*) \approx .37$, $B_2(\rho^*) \approx .01$



Simulation results for one lightpath (1 Gb/s)



- traffic mix
  - 20% backlogged +
  - 60% backlogged ×
  - 100% backlogged □
- saturation at load .65 due to transmitter blocking
- traffic mix
  - 0% backlogged +
  - 20% backlogged ×
  - 60% backlogged □
- negligible delay until saturation

Conclusions

- a generalized polling system for a passive optical MAN
  - building on EPON and TWIN
- our contributions:
  - a new asynchronous MAC protocol
  - flow-aware grant allocations for implicit service differentiation
  - excellent, predictable performance: flow throughput and packet latency
- extensions (work in progress):
  - sharing multipoint-to-multipoint lightpaths in a passive optical wide area network
  - application to data centres and the cloud