UCDAVIS TEKNOVUS NOKIA

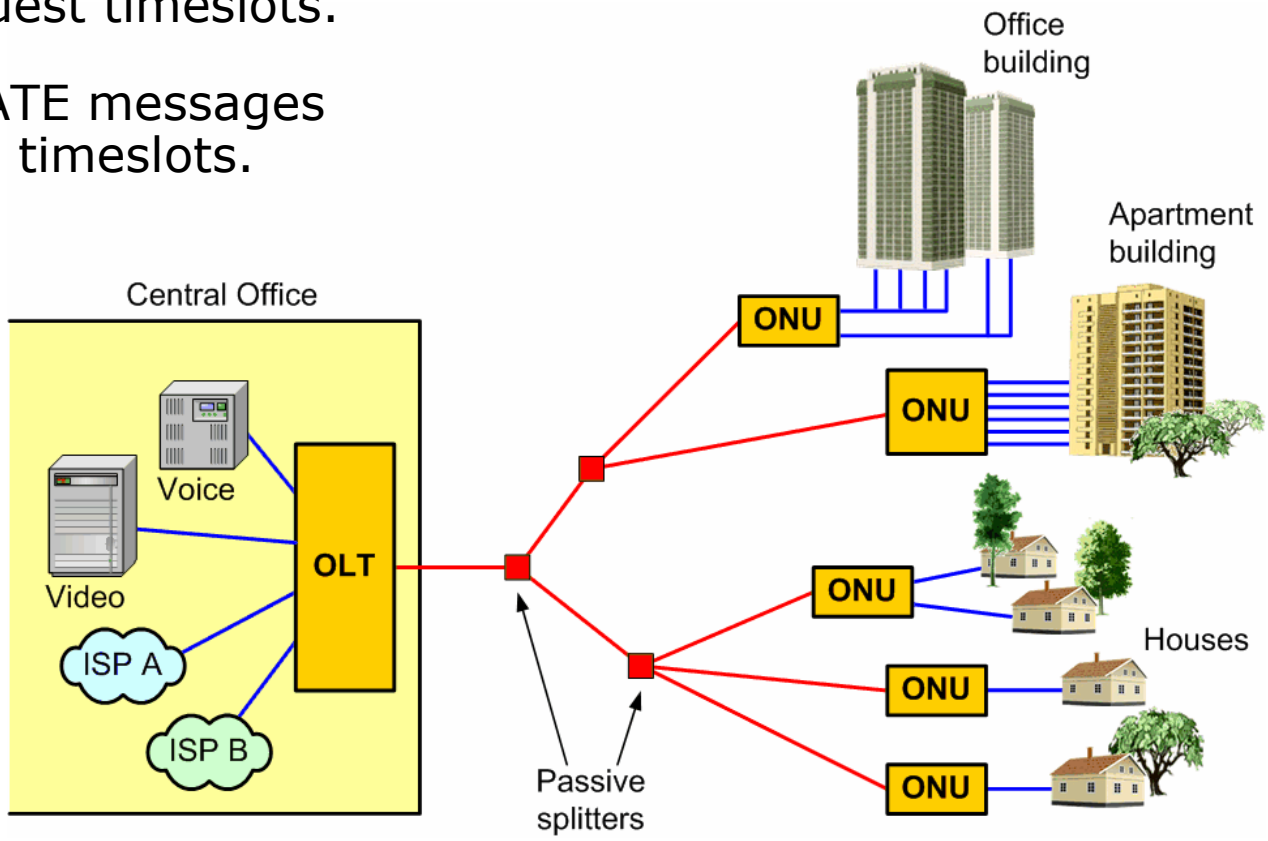
Fair Queuing with Service Envelopes (FQSE):

a Cousin-Fair Hierarchical Scheduler for Ethernet PON

Glen Kramer, Teknovus, Inc
Amitabha Banerjee, UC Davis
Narendra Singhal, UC Davis
Biswanath Mukherjee, UC Davis
Sudhir Dixit, Nokia Research Center
Yinghua Ye, Nokia Research Center

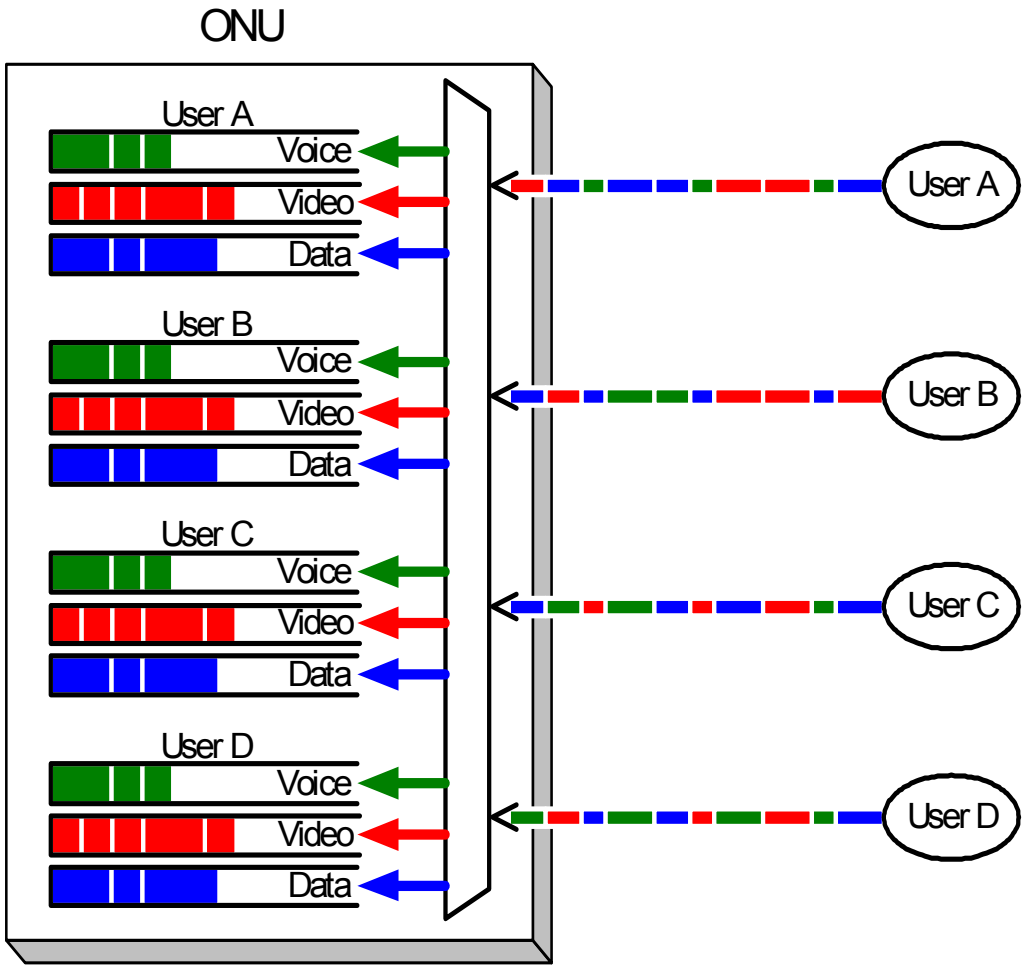
EPON Architecture

- All transmissions are performed between **Optical Line Terminal (OLT)** located in CO and **Optical Network Units (ONUs).**
- ONUs are granted time-shared access to the medium.
- ONUs send REPORT message to the OLT to request timeslots.
- The OLT sends GATE messages to ONUs to assign timeslots.



ONU's Configuration

- An ONU may contain multiple queues and serve multiple users.
- Each queue should get guaranteed bandwidth B^{MIN} regardless of the state of other queues.
- A queue should not be given more bandwidth (slot size) than it can use.



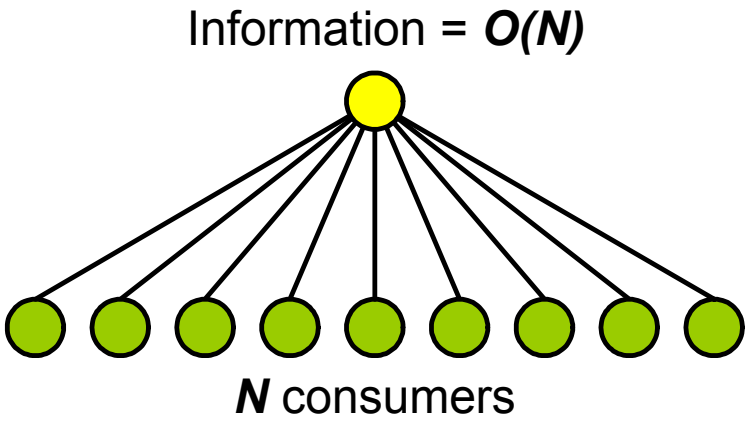
Fairness Requirement

- Excess bandwidth B^{EX} left by idle queues should be distributed between backlogged queues in a fair manner, i.e., in proportion to weight φ_i assigned to each queue i.

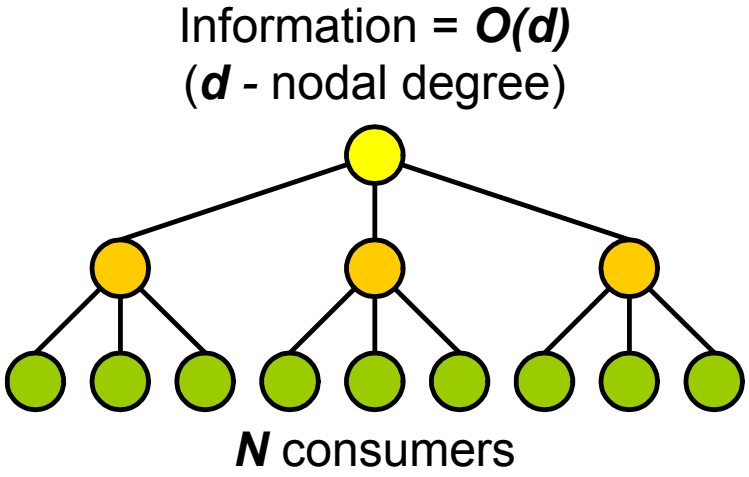
- The excess bandwidth should be distributed between the backlogged queues fairly regardless of whether the queues are located in the same ONU or in different ONUs.

Definitions (I)

- A **single-level-feedback scheduler** is a scheduler which receives status information from each consumer (Example: in IEEE 802.3ah the OLT receives queue sizes of each queue).

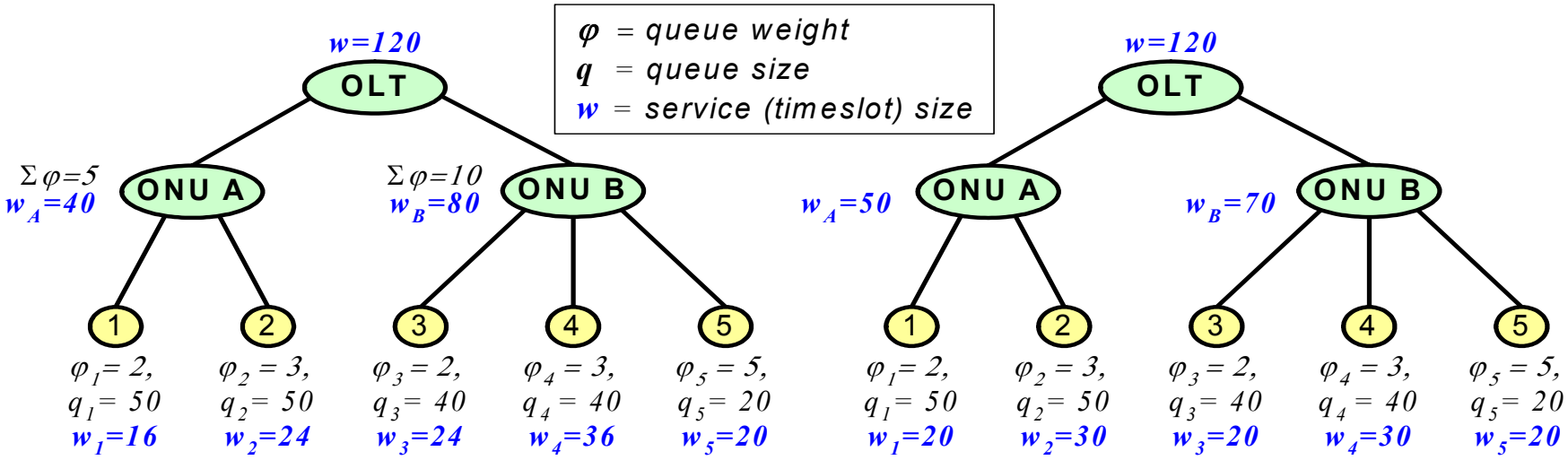


- A **hierarchical-feedback scheduler** is a multi-level scheduler where each node receives status information only from its immediate children.



Definitions (II)

- A hierarchical scheduling system is **sibling-fair** if each node distributes resources fairly among its immediate children.
- A hierarchical scheduling system is **cousin-fair** if resources are distributed such that each final consumer (leaf of a tree) receives a fair share.



(a) Sibling-fair (locally-fair) hierarchical scheduler

(a) Cousin-fair (globally-fair) hierarchical scheduler

Problem Statement

- Single-level schedulers are not scalable with the number of consumers (queues).
  - Example: 32 ONUs × 64 subscribers/ONU × 3 queues/subscriber = 6144 queues ≈ **52%** bandwidth consumed by REPORT messages.
- Hierarchical-feedback schedulers are scalable, but are not cousin-fair. Backlogged queues with identical SLAs located in different ONUs may get non-equal service.
- Our goal is to design a cousin-fair hierarchical-feedback scheduling algorithm.

FQSE

- FQSE is based on a concept of service envelope (SE).
- SE represents amount of service (timeslot size) given to a node as a function of some non-negative value called satisfiability parameter (SP).
- The meaning of SE function: as the SP changes, the SE determines the fair amount of service that should be given to a consumer (queue, ONU).
- Under heaviest load ($SP = 0$), exactly the guaranteed minimum timeslot will be given to the consumer, i.e., the consumer will get its guaranteed minimum service.
- As SP value increases, the consumer will be given an additional timeslot (surplus bandwidth) equal $\varphi_i \times SP$. When the timeslot size reaches $q_{i,k}$ (total queue length), it will not increase anymore, even if SP increases.

Construction of SE function

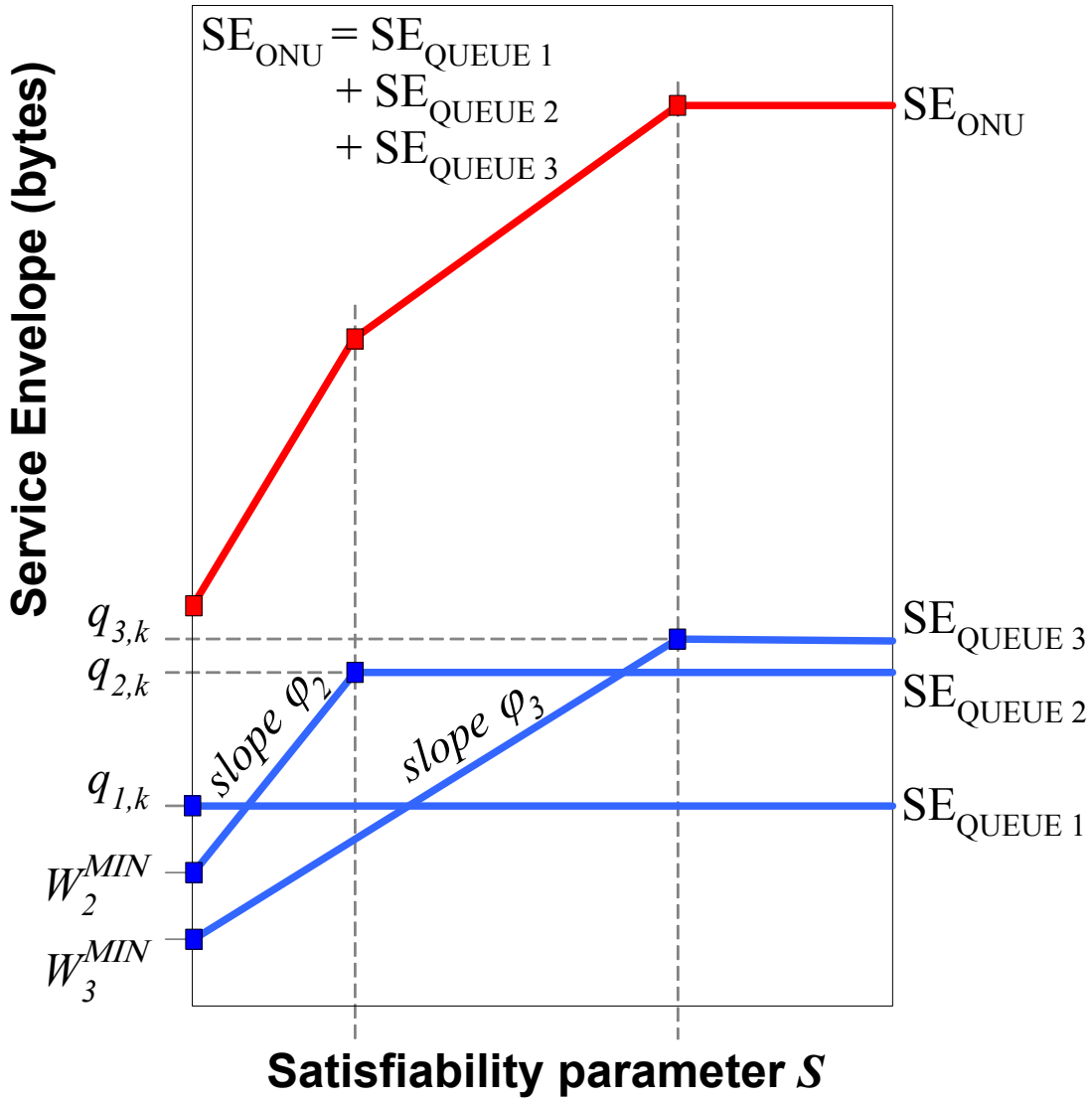
W_i^{MIN} – guaranteed slot of queue i

$q_{i,k}$ – length of queue i in cycle k

φ_i – weight of queue i

Typically SE_{QUEUE} consists of two linear segments (queues 2 and 3).

If $q_{i,k} < W_i^{MIN}$ SE_{QUEUE} consists of one segment (queue 1).



Phase 1 of FQSE: Requesting Service

- **Step 1:** ONU collects SE_{QUEUE} from all the queues in this ONU and builds SE_{ONU} ($SE_{ONU}=\sum SE_{QUEUE}$) which it must send to the OLT in a REPORT message. The function is to be transmitted as an array of point coordinates.

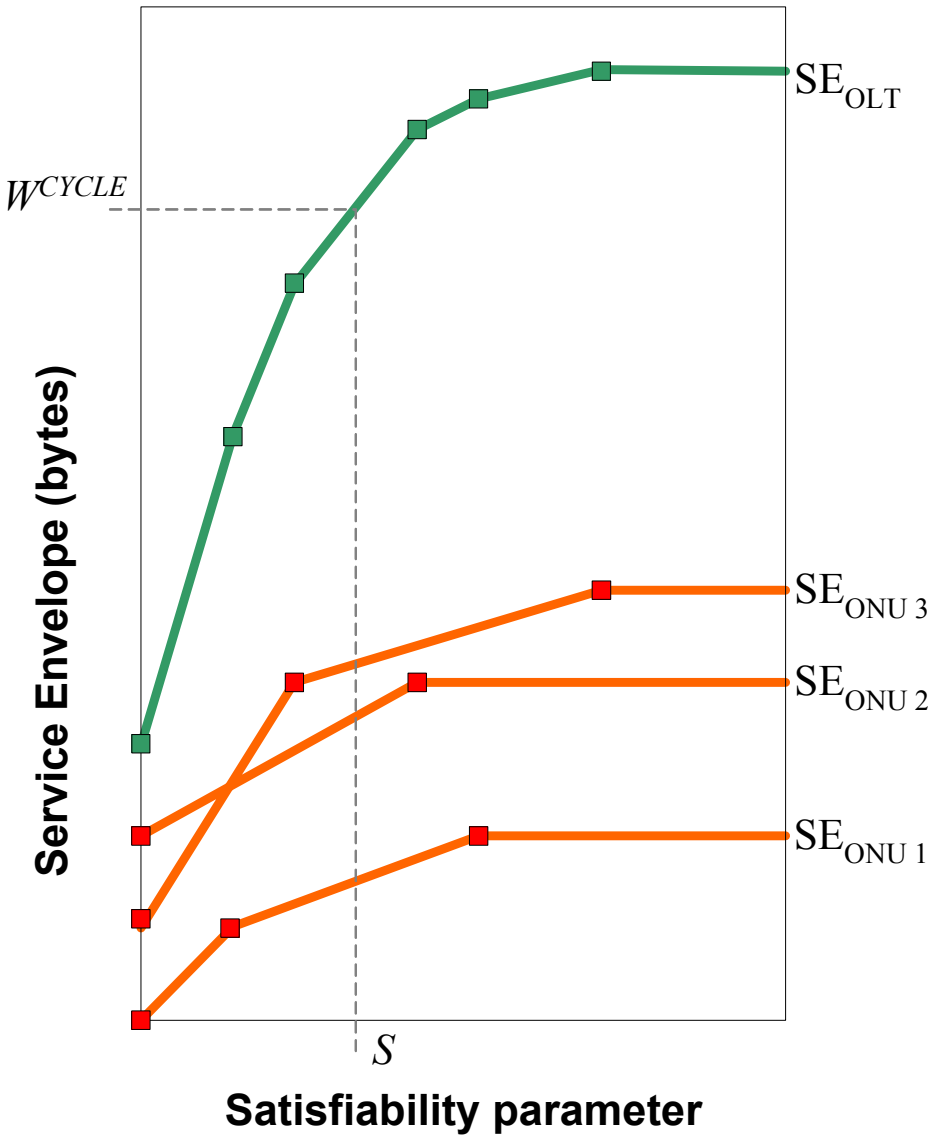
- **Step 2:** The REPORT message can accommodate only K points. If number of points in SE_{ONU} exceeds K, the ONU will perform a piecewise-linear approximation to describe SE_{ONU} with only K points. (This step makes it a hierarchical-feedback scheduler, as the amount of information sent to the OLT does not depend on the number of queues. anymore).

- **Step 3:** The OLT waits for the REPORT messages from all the ONUs. Once all the messages are collected, the OLT builds the SE_{OLT} function as $SE_{OLT}=\sum SE_{ONU}$.

Phase 2 of FSSE: Granting Service

- **Step 4:** The OLT knows the cycle capacity W^{CYCLE}. When the OLT obtains the SE_{OLT}, it calculates the SP S as $W^{CYCLE} = SE_{OLT}(S)$. The OLT then transmits the value S to each ONU in a GATE message.

- **Step 5:** Upon receiving the GATE message, an ONU assigns a slot of size $w_{i,k}=SE_{QUEUE}(S)$ to each queue i in that ONU. This assigns each queue a fair slot size.

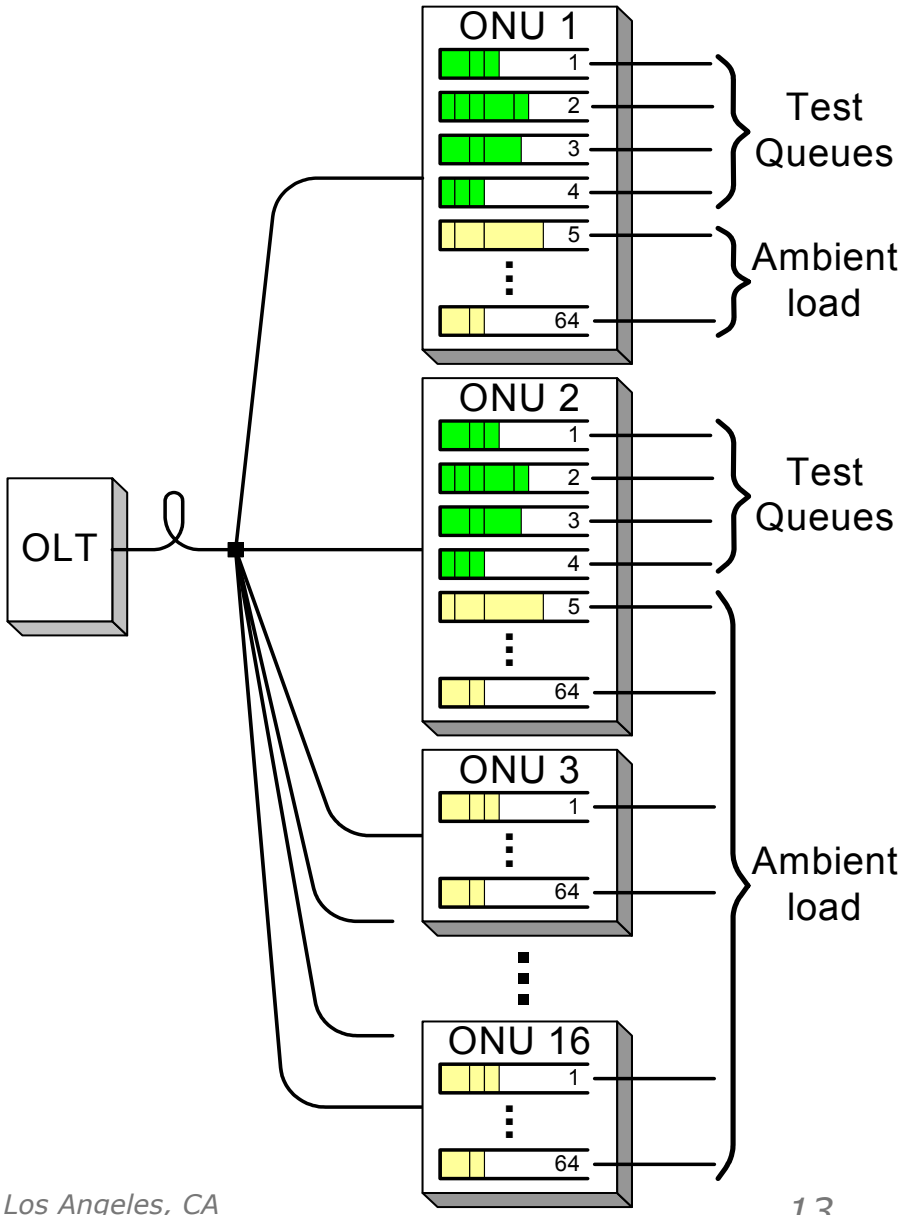


Optimizations for Packet-based Networks

- The above 5 steps represent a foundation of FQSE.
  - Suitable for raw data or ATM (service quanta are small and constant).
- But Ethernet consists of indivisible variable-sized packets. Need to solve 2 problems:
  - Head-of-Line blocking
    - Happens when a queue is given a small slot (to achieve fairness), but has a larger packet at the head of the queue.
    - Solved by allowing queue to overdraw on its service and compensate for it later.
  - Bandwidth (timeslot) utilization
    - Happens because variable-size packets don't fill the timeslot completely.
    - Solved by (1) combining all unused slot remainders and (2) maintaining deficit counters per queue. A queue with the largest deficit will be allowed to use the remainder to reduce its deficit.

Simulation Experiment

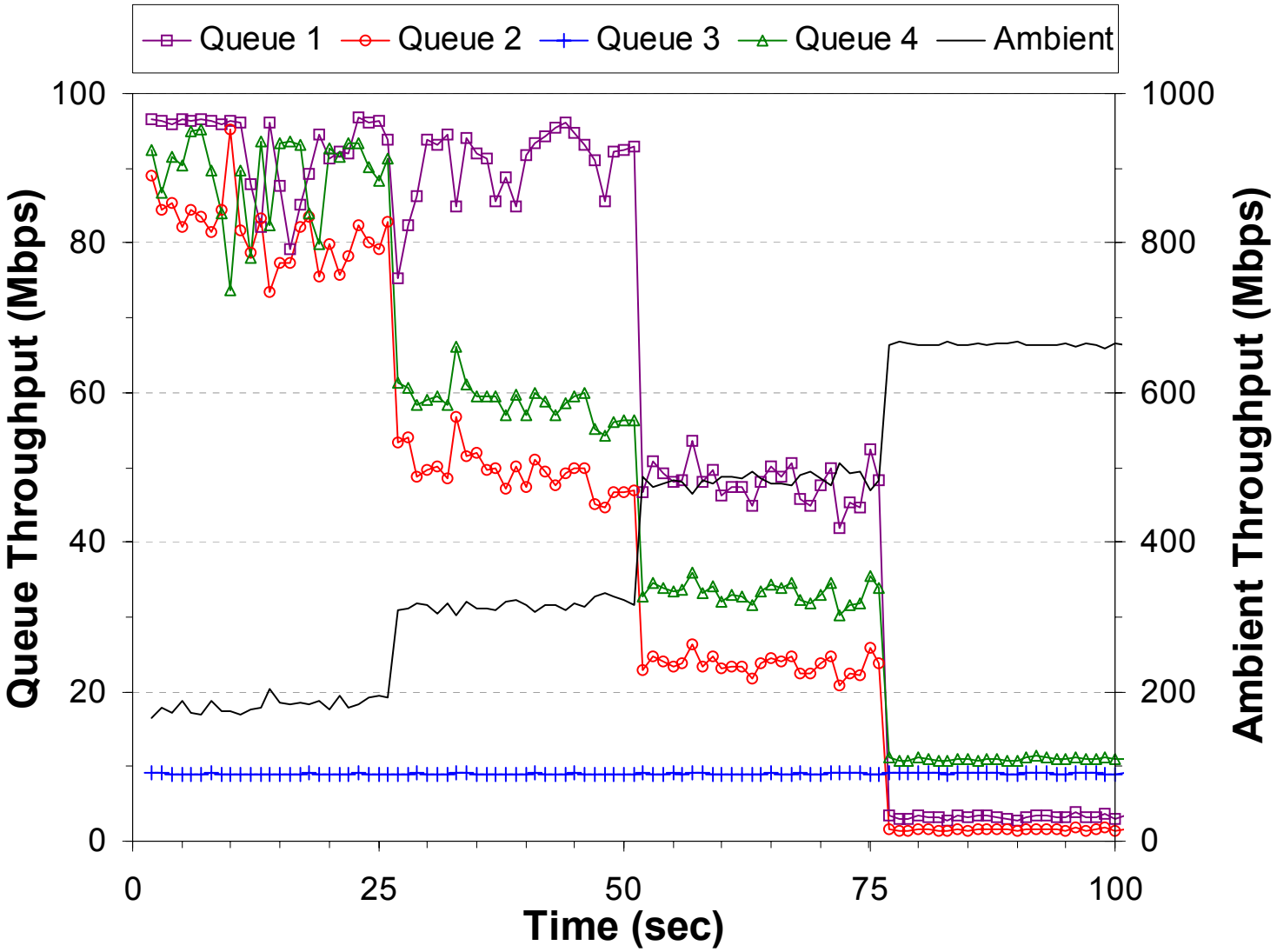
- 16 ONUs
- 64 queues per ONU
- Measure throughput of 4 test queues in two different ONUs as the ambient network load changes.
- Use self-similar traffic.
- Run for 100 seconds, then increase ambient load. Repeat 4 times.



Types of Test Queues

Queue	B_{MIN}	φ	Description
1	0	2	Best-effort service (no guaranteed bandwidth). Under very heavy load may get no service.
2	0	1	Good-effort service. If excess bandwidth is available, this queue will get half the bandwidth that the queue 1 gets.
3	10 Mbps	0	Fixed-bandwidth service (no extra bandwidth). Used for circuit-emulation services.
4	10 Mbps	1	Always gets its guaranteed bandwidth (10 Mbps) plus excess bandwidth if the network load is not high.

Throughput Analysis



Throughput Analysis (cont.)

- While the ambient load is low (interval 0 – 100s), each test queue can be served completely (to exhaustion). Throughput is bursty reflecting bursty arrivals.
- As the ambient load increases, the test queues remain backlogged between timeslots.
- When the queues are backlogged (i.e., bandwidth is scarce), queue's throughput should approach its fair value.

Fairness Bound

- Serving variable-sized indivisible packets introduces some inaccuracy (unfairness).

- **Theorem:** Service w_i' received by queue i in any interval of n cycles (n = 1, 2, 3,…), during which the queue remains backlogged, does not deviate from the optimal (fair-share) service w_i by more than S^{MAX}-1:

$$|w_i' - w_i| < S^{MAX}$$

Conclusion

- FQSE is hierarchical-feedback scheduler
  - Amount of work at each node is bounded by $O(D \times \log D)$ (D = nodal degree).
- FQSE is cousin-fair
  - Fairness error over any interval is bounded by maximum packet size S^{MAX}.

For complete algorithm details and pseudo-code see
G. Kramer, A. Banerjee, N. Singhal, B. Mukherjee, S. Dixit, Y. Ye, "*Fair Queuing with Service Envelopes (FQSE): a Cousin-Fair Hierarchical Scheduler for Subscriber Access Networks*", to be published in JSAC in Q3, 2004.