The University of Hong Kong

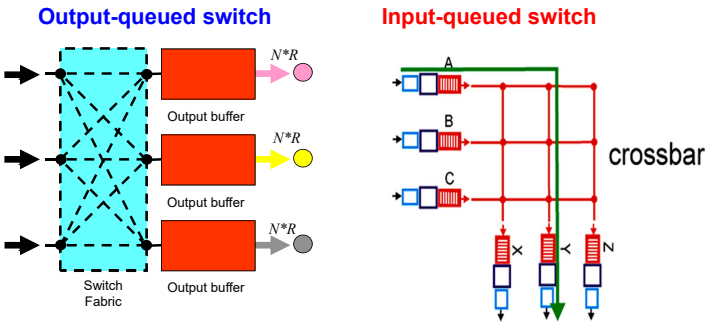
Designing the Most Efficient Iterative Scheduling Algorithms for Input-queued Switches

Lawrence Yeung
Department of Electrical & Electronic Engineering
The University of Hong Kong
Dec. 1, 2016

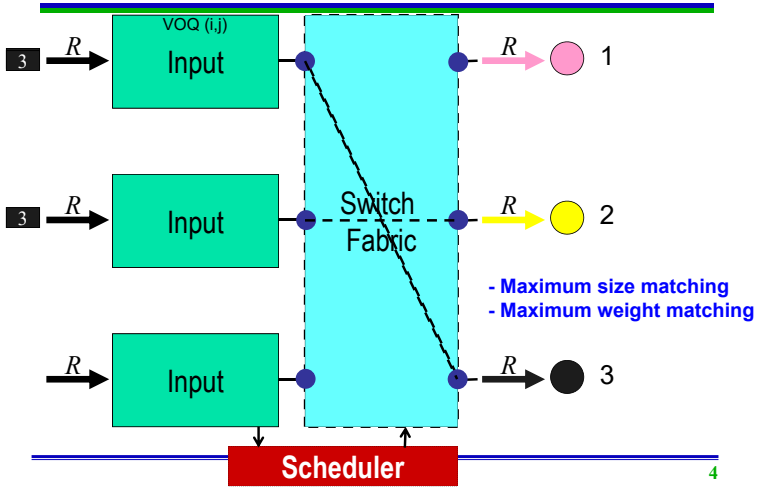
Outline

- Background
  - Input queued switch
  - Iterative scheduling algorithms
- Highest rank first (HRF)
  - HRF-basic
  - HRF-refined
  - HRF with request coding (HRF-RC)
- Performance evaluations
- Conclusions

Two Types of Switches



Input-queued Switch



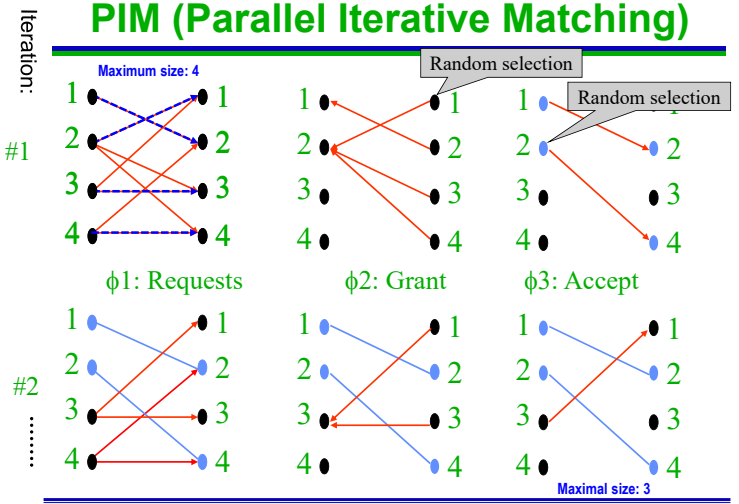
Iterative Scheduling Algorithms

- Maximal size matching (MSM) is simpler
  - as no backtracking on established connections.
- Iterative scheduling algorithms are good for finding MSM, and hardware implementation.
- Each iteration consists of 3 phases:
  - Request: Inputs send matching requests to outputs
  - Grant: Each output grants at most one request
  - Accept: Each input accepts at most one grant

1) An iterative MSM algorithm guarantees maximal size matching in *N* iterations, where *N* is the switch size.
2) In practice, only a small fixed number of iterations are used.

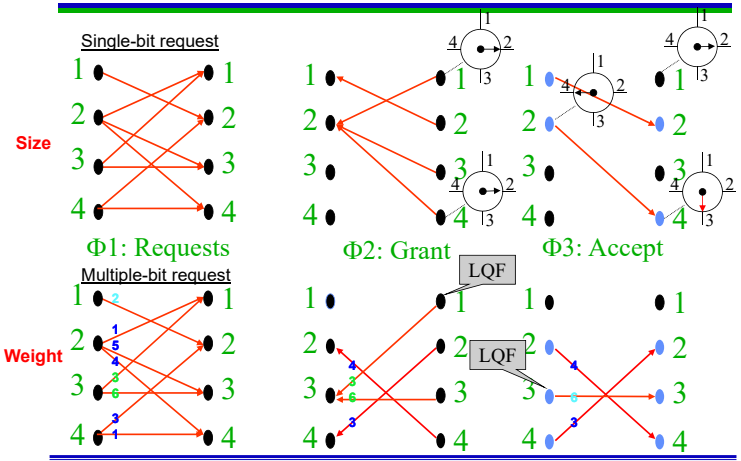
Request: only for VOQ > 0
Grant/accept: only for winners

PIM (Parallel Iterative Matching)



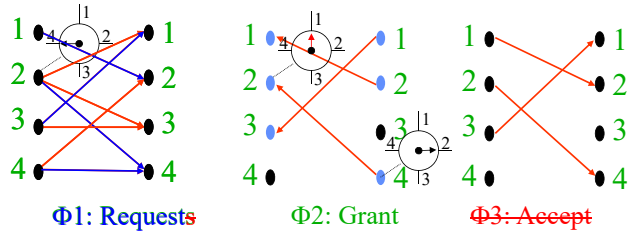
A matching is of maximal size if "no input or output is left unnecessarily idle".

*i*SLIP (iterative RR with slip)



***i*LQF**

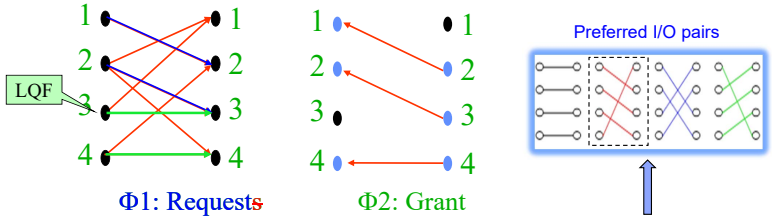
DRRM (Dual RR Matching*)



- Single request from each input
  - Not to unnecessarily attract > 1 grants (but ..)
  - A grant is guaranteed to be accepted => 2-phase, simpler
- Single-iteration performance comparable to iSLIP-1

* Yihan Li, Shivendra Panwar and H. Jonathan Chao, "On the Performance of a Dual Round-Robin Switch," IEEE INFOCOM 2001, vol. 3, pp. 1688-1697, April 2001

SRR (Synchronous RR*)



- Single request from each input based on a global RR (gRR) schedule.
  - Implicit; no local RR arbiters, simpler
- Scheduling priority is given to
  - preferred I/O pair first, and longest VOQ next.
- Outperforms iSLIP-1 & DRRM under uniform traffic

* A. Scicchitano, A. Bianco, P. Giaccone, E. Leonardi and E. Schiattarella, "Distributed scheduling in input queued switches" *IEEE ICC 2007*, June 2007, Glasgow, Scotland.

Iterative Scheduling Algorithms

Non-weighted matching:
- iSLIP* / DRRM / ...
- Rotating priority via local RR arbiters
- TDM-like high-load performance

Weighted matching:
- iLQF* / …
- Queue-based priority, where the LQF is always served first
- But difficult to implement, and *size* is limited

Hybrid:
- SRR ≈ gRR (*size*) + LQF (*weight*)
- What is the right balance between *size* and *weight*?

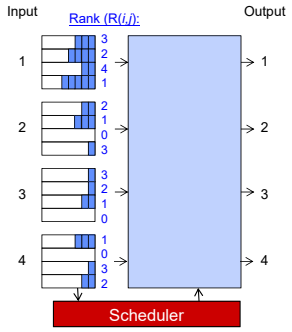
Our goal: A single-iteration scheduling algorithm that is simple to implement and better in performance.

(A minor change can have a big impact on performance!)

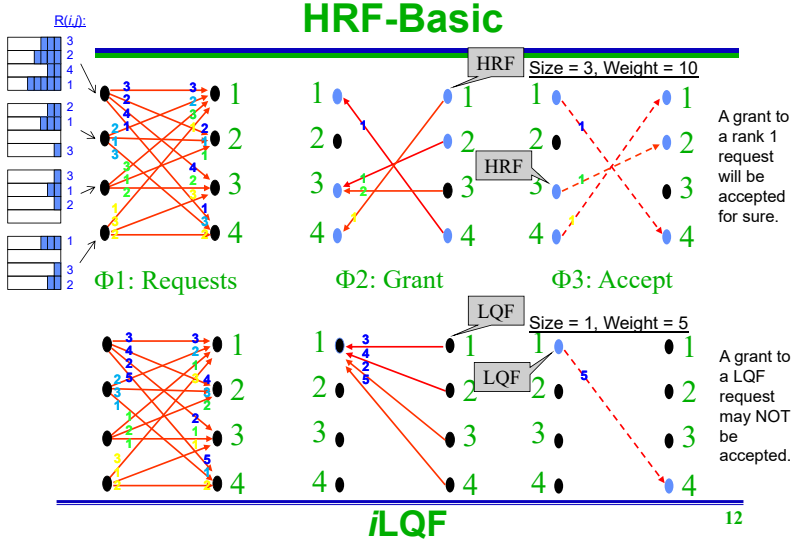
* N. McKeown, "Scheduling algorithms for input-queued cell switches," *PhD. Thesis*, University of California at Berkeley, 1995.

Rank-based Priority: HRF

- Each input ranks its *N* VOQs according to queue size.
  - *N* ranks (1 to *N*)
  - A special rank, R(*i*,*j*) = 0, is reserved for empty VOQ
    - ➔ log (*N*+1) bits
  - In arbitration, priority is given to VOQ with the highest rank, i.e. HRF
  - Rank-based priority vs queue-based priority



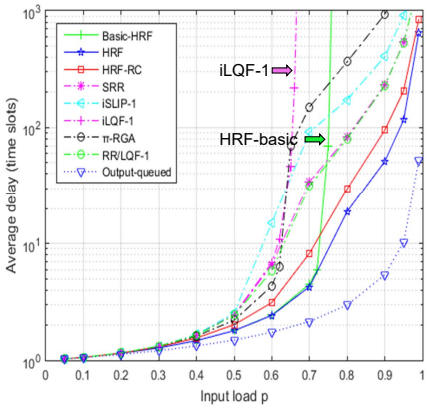
HRF-Basic



***i*LQF**

E.g. under uniform traffic

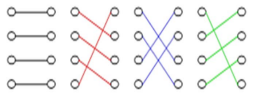
- HRF-basic vs iLQF-1
  - Rank-based priority is more effective
- BUT
  - Poor high-load performance
  - *Multiple-bit requests*



HRF-Refined

- gRR (as in SRR):
  - Each input has a distinct preferred output in each slot.
  - Each input *prefers* each output exactly once in every *N* slots.
  - Input *i* at time slot *t*, its preferred output *j* is given by

$$j = (i + t) \bmod N$$



- Scheduling priority is given to
  - preferred input-output pair first, and
  - highest rank VOQ next.

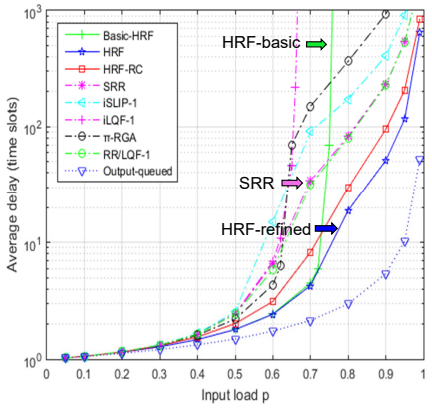
HRF-Refined

- Request: If output *j* is the preferred output and VOQ(*i,j*) > 0, input *i* sends 1 to output *j* and 0 to all others. Otherwise, send R(*i,j*) to all.
- Grant: An output grants the request from its preferred input first. If no preferred request, grants the request with the highest rank.
- Accept: Input accepts the grant from its preferred output. If no preferred grant, accepts the grant with the highest rank.

Note: Rank 0 = “empty”

E.g. under uniform traffic

- HRF-refined vs HRF-basic
  - High-load performance is improved
- HRF-refined vs SRR
  - HRF + gRR
  - LQF + gRR

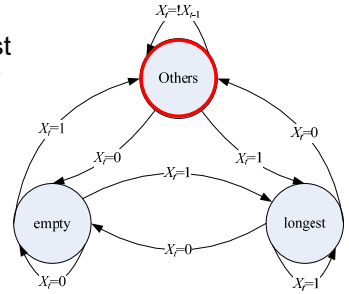


HRF with Request Coding (HRF-RC)

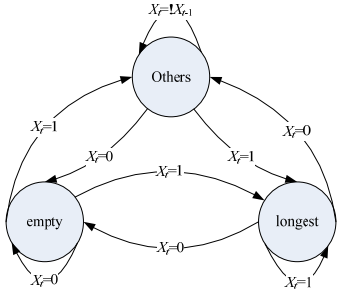
- Multiple-bit request ➔ single-bit request

- Idea: use the single-bit request (X_t) to indicate the increase or decrease of the VOQ rank
  - vs "empty" or "non-empty"

- Maintaining full-rank info at each input?
  - HRF-basic: successful VOQs ranked high
  - Our approach: **3 ranks**



Request Coding & Decoding



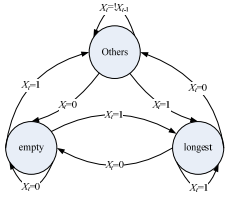
- Based on the value of $X_t X_{t-1}$

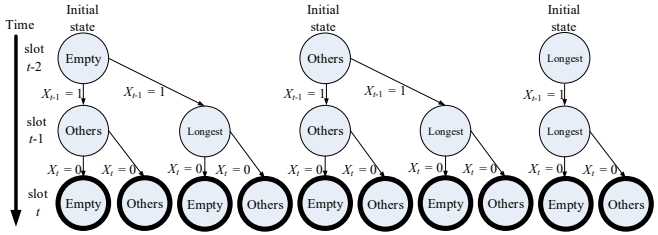
$X_t X_{t-1}$,	0 0	0 1	1 0	1 1
Ranks	Empty	Empty Others	Others Longest	Longest

Priority: Lowest ----------> Highest

E.g. $X_t X_{t-1}$="01"

- All possible state changes for $X_t X_{t-1}$="01"





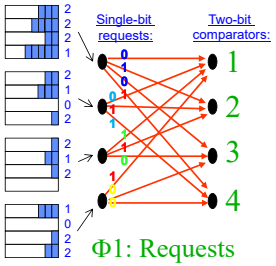
$X_t X_{t-1}$,	0 0	0 1	1 0	1 1
Ranks	Empty	Empty Others	Others Longest	Longest

HRF-RC

- Request: If an input's preferred output is backlogged at slot *t*, sends X_t = 1 to output *j* and X_t = 0 to others. Otherwise, using the original RC.
- Grant: Each output decodes X_t from
  - its preferred input as an occupancy indicator (VOQ(*i*,*j*) = 0 or not), and
  - other inputs using the $X_{t+1}X_t$ decoding table
- Accept: Each input accepts the grant from its preferred output first. Otherwise, accept the grant with the highest rank.

Properties of HRF-RC

- Simple to implement:
  - Three VOQ states/ranks
  - Single-bit request
  - Two-bit comparators



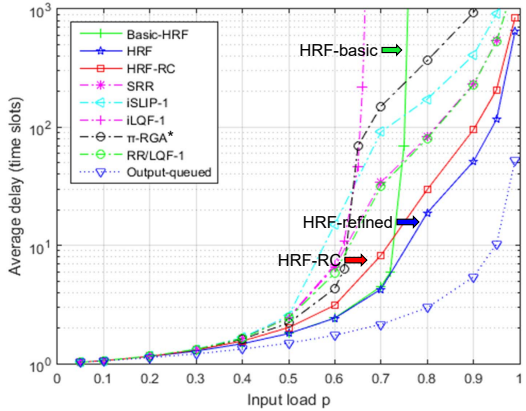
- HRF-RC is stable if each flow's arrival rate ≤ 1/*N*.
  - iSLIP & DRRM are stable under *uniform* traffic (≤ 1/*N*).
- HRF-RC satisfies the max-min fairness criteria.
  - iSLIP & DRRM ensures no starvation.

Outline

- Iterative scheduling algorithms
- Highest rank first (HRF)
  - HRF-basic
  - HRF-refined
  - HRF with request coding (HRF-RC)
- **Performance evaluations**
- Conclusions

Uniform

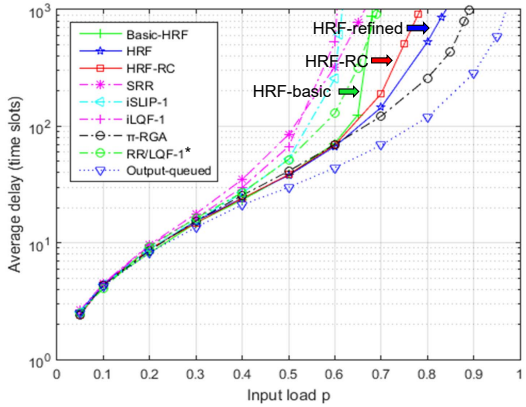
- 64 x 64 switch



* S. Mneimneh, "Match form the first iteration: an iterative switching algorithm for input queued switch," *IEEE/ACM Trans. on Networking*, Vol. 16, Issue 1, pp. 206 – 217, Feb. 2008.

Bursty

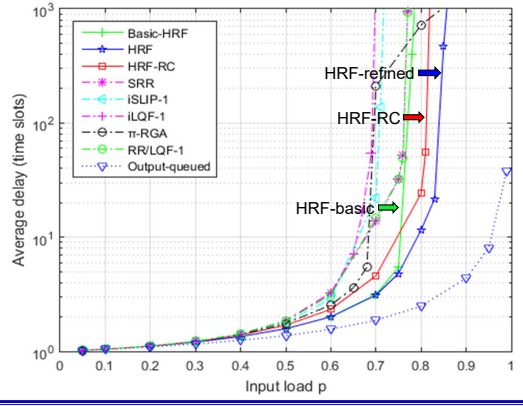
- Burst size = 30 cells



* B. Hu, K. L. Yeung, Q. Zhou and C. He, "On Iterative Scheduling for Input-queued Switches with a Speedup of 2-1/*N*," Accepted by *IEEE/ACM Transactions on Networking*, Feb. 2016.

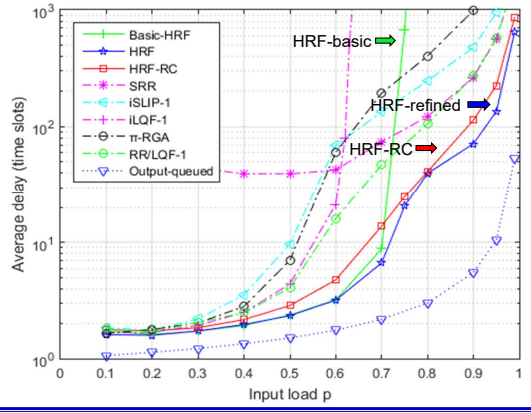
“Output” Hotspot

- Each input has a distinct hotspot output.



“Input” Hotspot

- Input 1 is always fully loaded.



Conclusions

- We reviewed existing work on iterative scheduling algorithm design.
- We proposed a rank-based priority scheme (HRF)
- We designed a request coding scheme for keeping single-bit request