A Low Power Approach to System Level Pipelined Interconnect Design

Vikas Chandra

Anthony Xu

Herman Schmit

Electrical & Computer Engineering
Carnegie Mellon University

Introduction

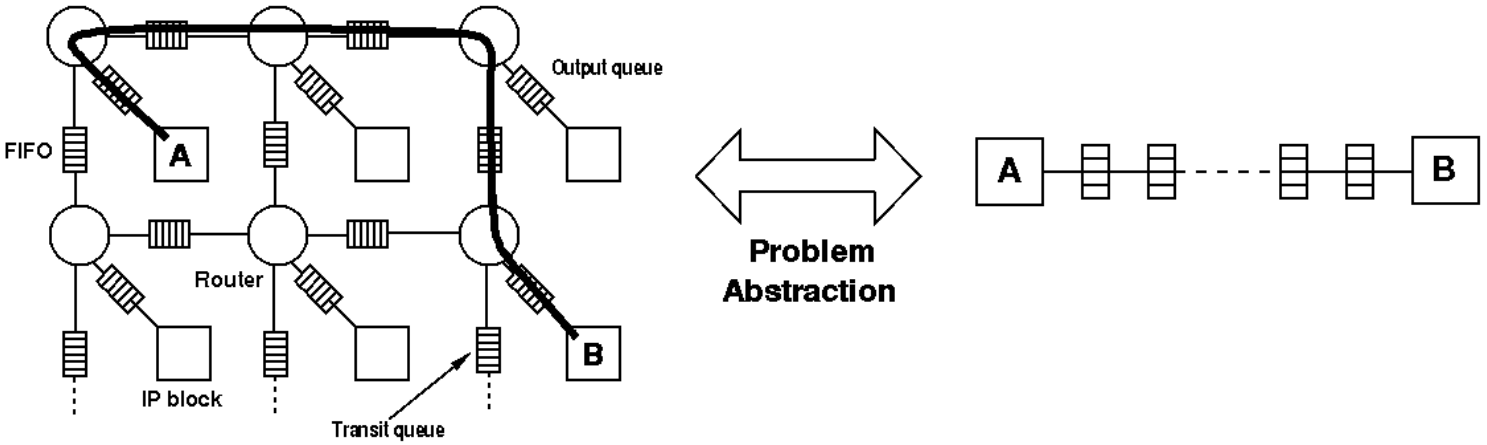
- **On chip communication issues**
  - Bottleneck for high performance designs
  - Bandwidth limited by on-chip busses
  - Ratio of global interconnect delay to average clock period continues to grow

- **Solutions**
  - On-chip network
    - Scalable interconnect bandwidth
  - Interconnect pipelining
    - Latch repeaters in wires

On-chip Network

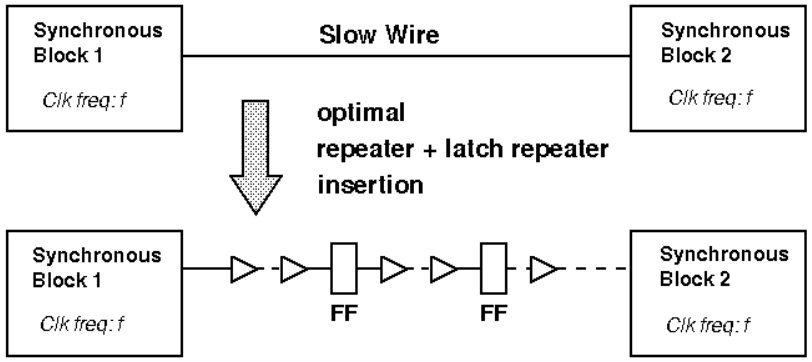
- **Network-on-Chip (NoC) provide**
  - Scalable interconnect bandwidth
  - Structured pipelining and re-buffering

- **Global interconnects in an NoC require**
  - buffering and flow-control

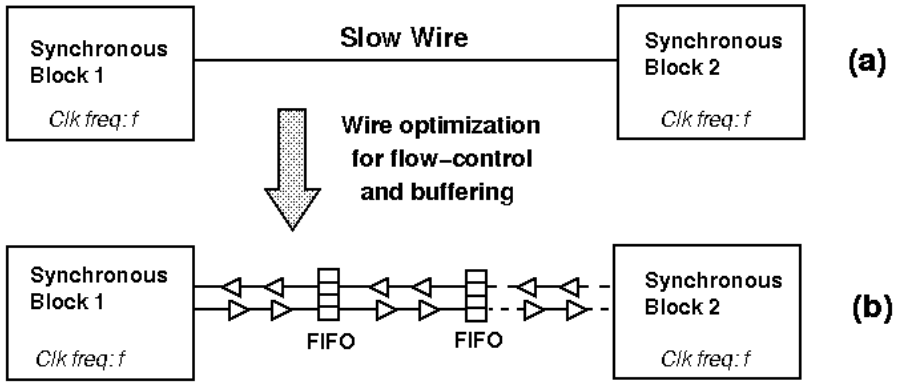


ASIC

- Wire optimization in an ASIC

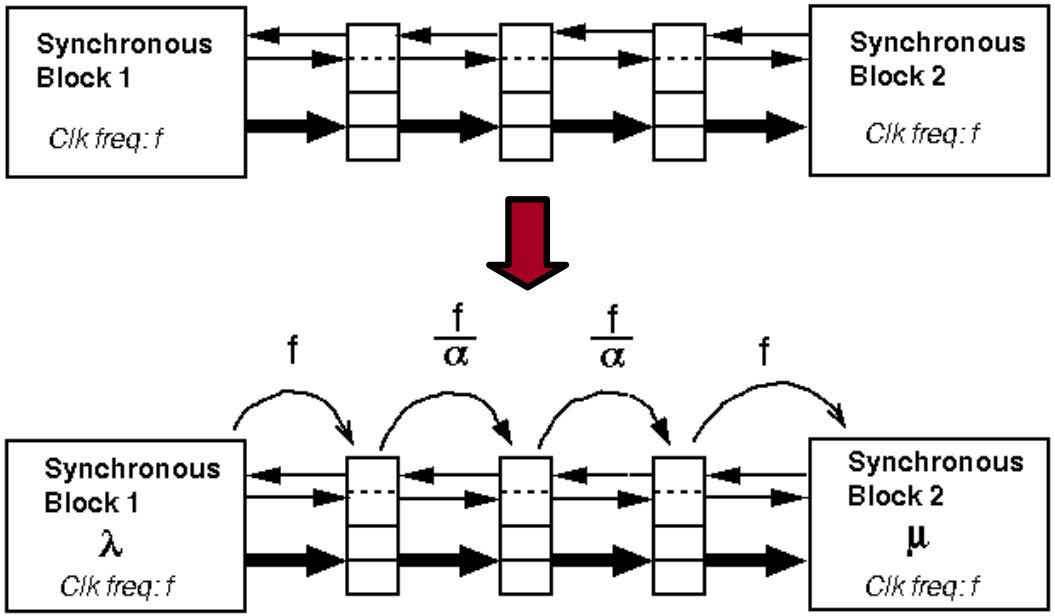


- Wire optimization in an ASIC having
  - Multiple clock domain, bursty data behavior



Problem Definition

- **Power-optimization of FIFO-based on-chip communication links**
  - Clock frequency scaling
  - Voltage scaling
  - Sizing of FIFOs in the communication links



Background - Terms

- **FIFO:**
  - *First-In-First-Out queue*
  - *Synchronous FIFO are also called elastic buffers*
- **FIFO size:**
  - *Maximum number of data elements in a FIFO*
- **Channel:**
  - *One or more FIFO connected in series*
- **Stage:**
  - *Each of the FIFOs in the channel is called a stage*

Assumptions

- **Synchronous system**
  - Source IP and destination IP run at the same clock frequency
  - FIFO voltages and clock frequencies are scaled down to save power
- **Discrete time system**
  - All events occur at positive edge of the clock
- **No data dropping**
  - Data in a FIFO waits if the next FIFO in the channel is full

System Parameters

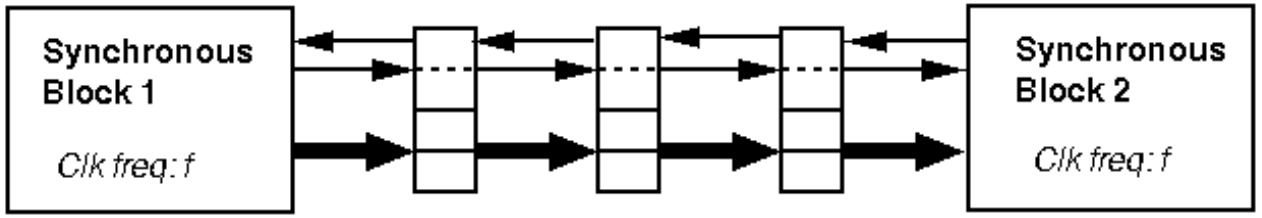
- **Queue behavior will depend upon**
  - Relative data production and consumption rates

- **l : Average data production rate**
- **m : Average data consumption rate**

- **Throughput requirement**
  - Number of data items read per time unit
  - Performance metric of the channel

Channel Analysis

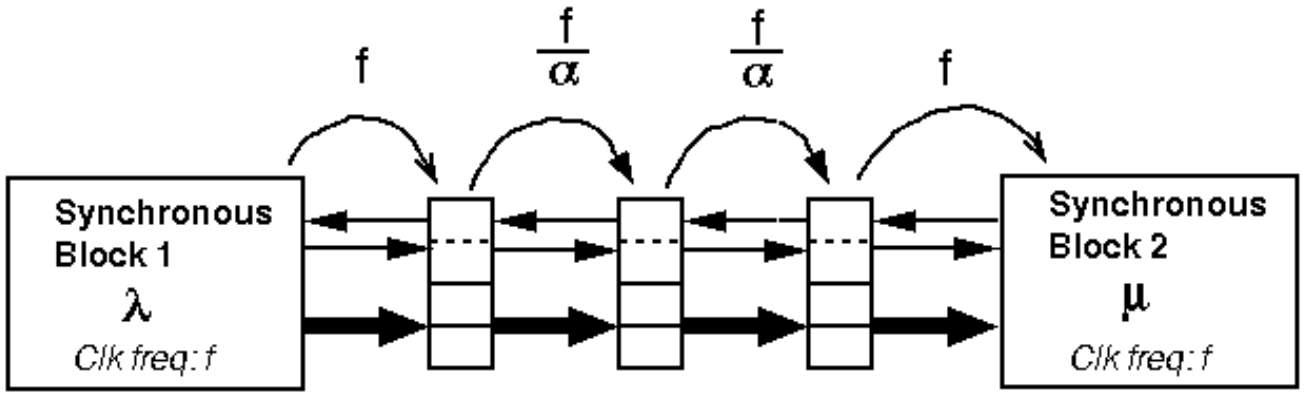
- **Designed an interconnect channel using an ASIC methodology**
  - Length of channel given by physical design information
  - Maximum frequency requirement constraint by the designer
  - Optimization results in 3 stage of latch repeaters
  - Replaced the latch repeaters with FIFOs to enable
    - Data bursts, Flow control



- The FIFOs are sized to meet the throughput requirement

Problem – revisited

- **Power aware FIFO sizing methodology**
  - Choice of power optimal frequencies and voltages for the FIFO stages
  - Resizing of the FIFOs in the channel to recover the performance
    - Performance loss due to clock scaling in the FIFO
  - Relation between λ and μ and the clock scaling factor of the FIFO



Power Optimization

- Dynamic power consumption in CMOS

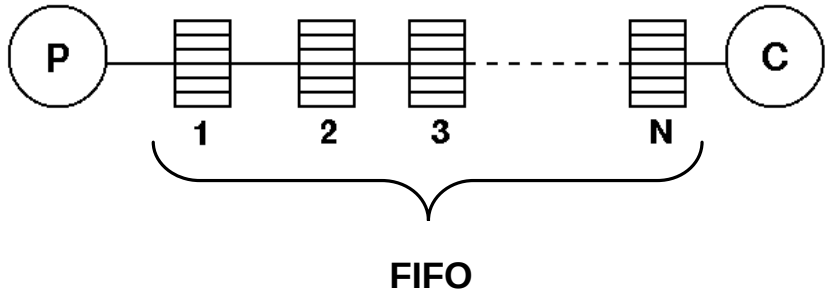
$$P_{dyn} = A_f \cdot C \cdot V^2 \cdot f$$

- Frequency and voltage are also related

$$f \propto \frac{(V_{dd} - V_t)^\gamma}{V_{dd}}$$

- Voltage and frequency island in the channel

Transfer Blocking



- **Can be formulated as a queueing network**
- **The queues in this work are M/M/1/K queues**
  - Difficult problem – no closed form solution
- **Analysis further complicated by**
  - **Transfer blocking** – Data waits if the next FIFO is full
- **Blocking increases with clock scaling**
  - Slow movement of data in the queue

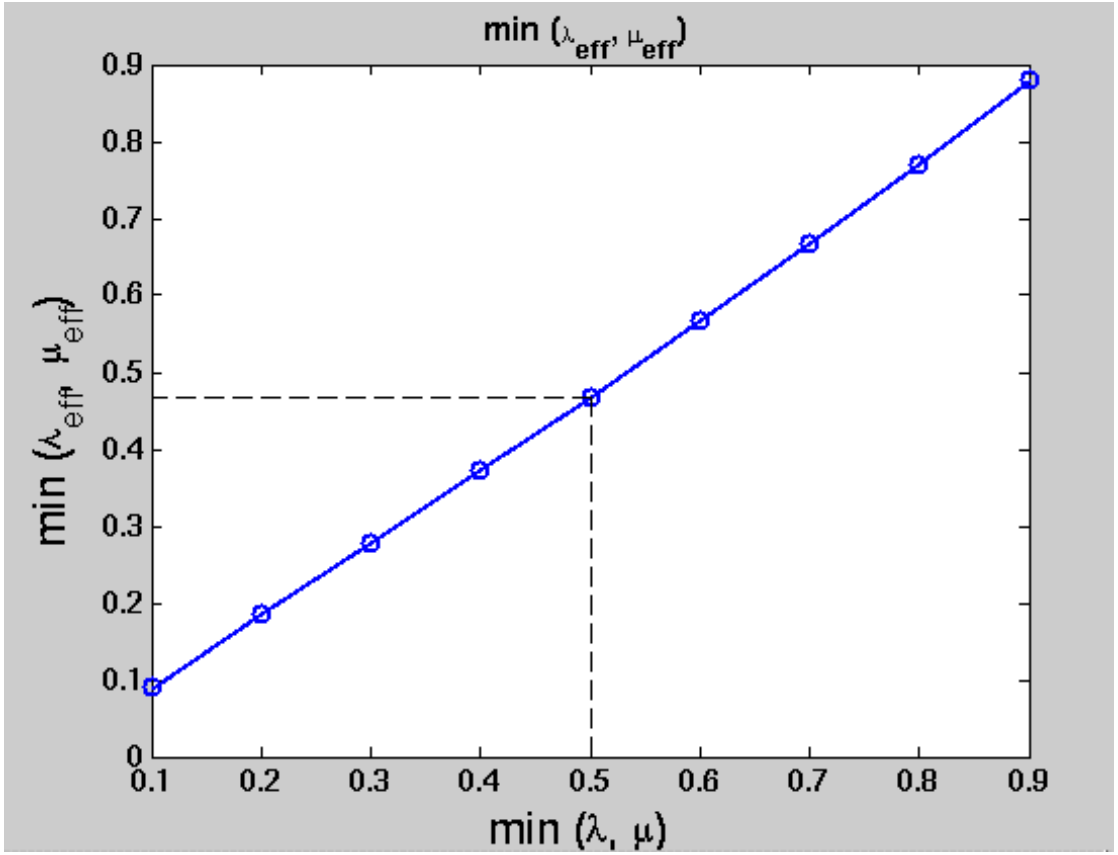
Effective l and m

$$\text{Channel data rate} = min\{(\lambda \cdot f), (\mu \cdot f)\}$$

- l_{eff}: effective value of l
- m_{eff}: effective value of m
- l_{eff} and m_{eff} are obtained by simulation
  - C simulator models the channel

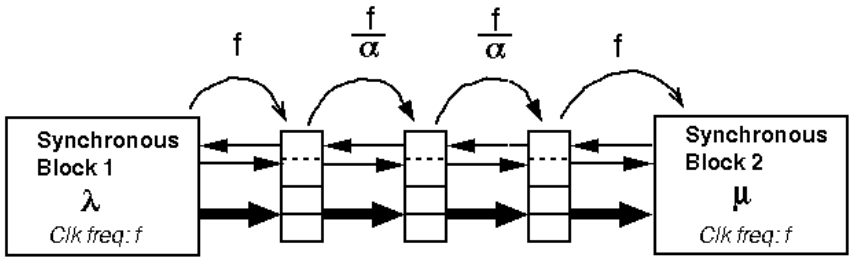
$$\begin{aligned} \text{Effective channel data rate} &= min\{(\lambda_{eff} \cdot f), (\mu_{eff} \cdot f)\} \\ &\leq min\{(\lambda \cdot f), (\mu \cdot f)\} \end{aligned}$$

Effective l and m



$$\begin{aligned} \text{Effective channel data rate} &= min\{(\lambda_{eff} \cdot f), (\mu_{eff} \cdot f)\} \\ &\leq min\{(\lambda \cdot f), (\mu \cdot f)\} \end{aligned}$$

Limits on frequency scaling



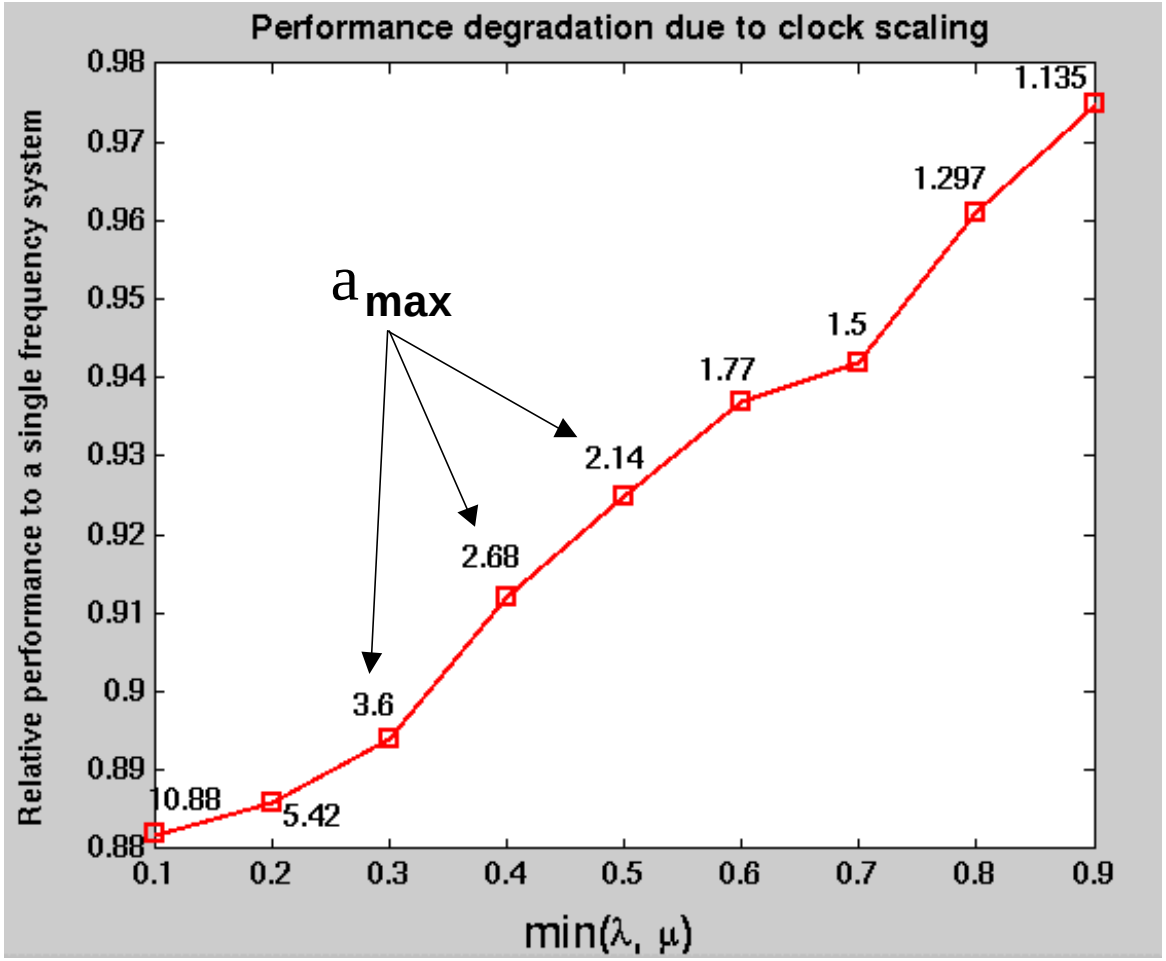
- **For a frequency scaled channel**
  - With no performance degradation

$$\alpha_{max} = \frac{1}{min(\lambda_{eff}, \mu_{eff})}$$

Theorectical upper bound on a

- **Channel frequency can be lowered to meet the effective channel data rate**
- **The maximum possible value of** a **is less than** a_{max}
  - Attributed to transfer blocking

Effect of transfer blocking



- Performance degradation proportional to a_{max}

Throughput – α tradeoff

- For each min(λ,μ), α is bounded

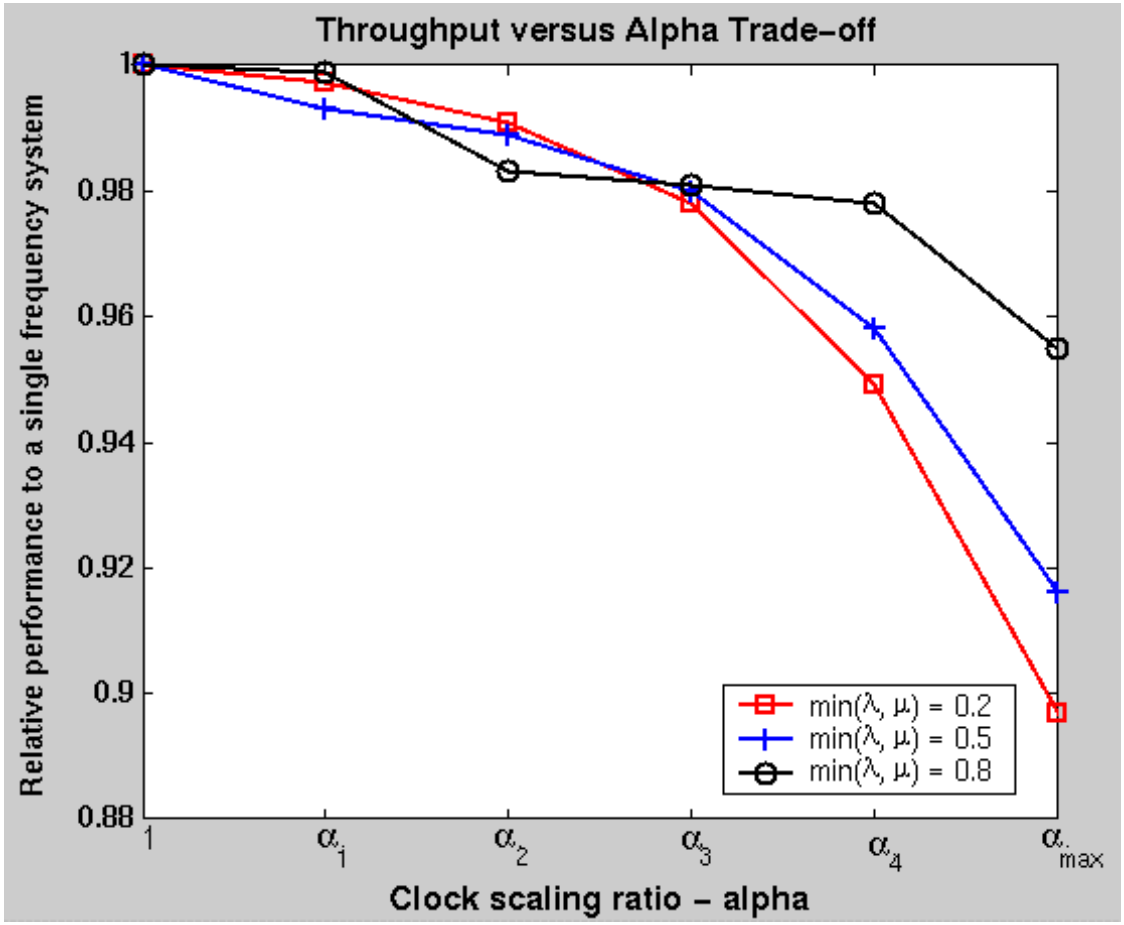
$$1 \leq \alpha \leq \alpha_{max}$$

- Four values of α chosen between 1 and α_{max}

$$\alpha_i = 1 + \left(\frac{\alpha_{max} - 1}{n} \right) \cdot i$$

- Three sets of system parameters were chosen
  - min(λ,μ) = 0.2
  - min(λ,μ) = 0.5
  - min(λ,μ) = 0.8

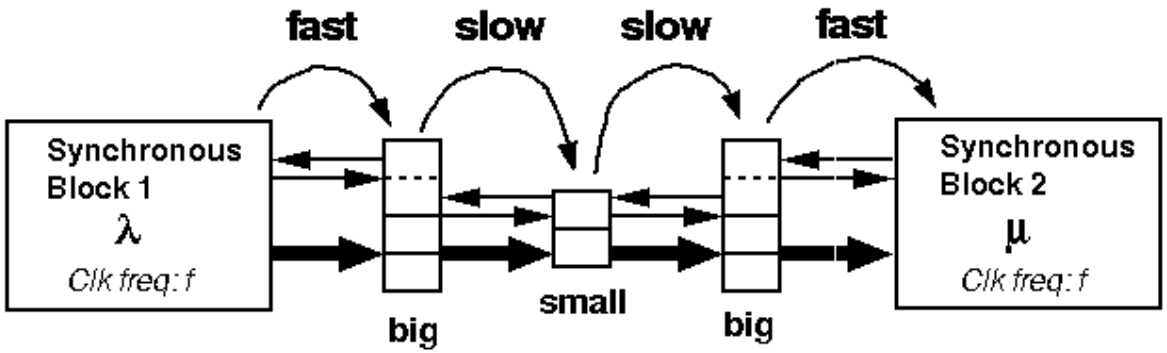
Throughput – a tradeoff



Throughput falls sharply at f/a_{max}

Performance recovery: FIFO resizing

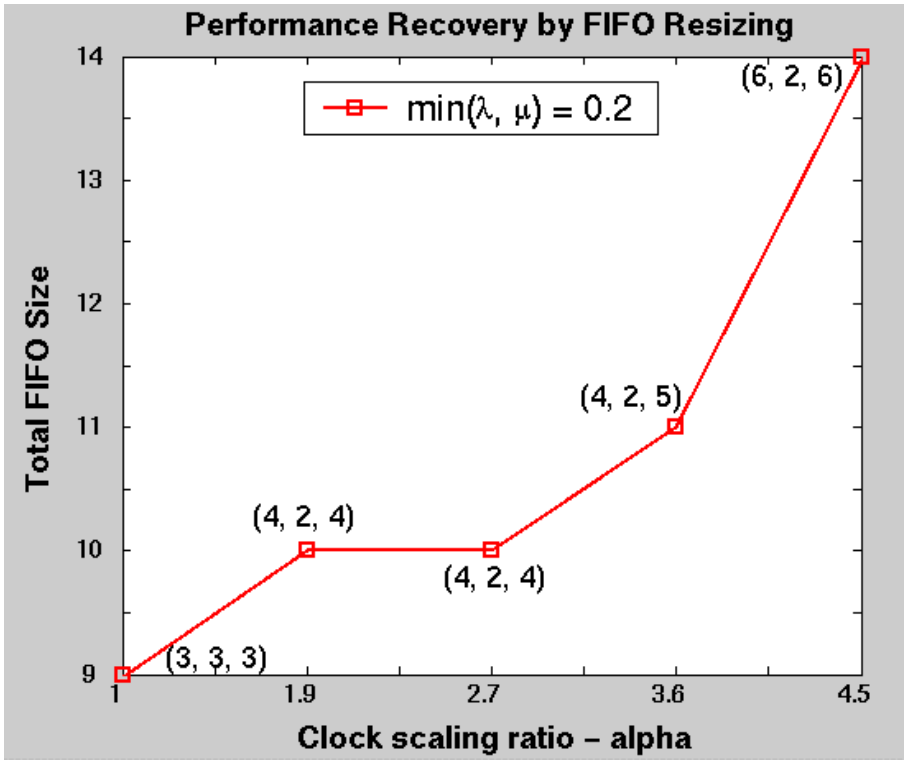
- **Increasing FIFO sizes improves performance**
  - Decrease in the occurrence of transfer-blocking
- **Write and read rates are different for beginning and end FIFOs**
  - Increasing their sizes have maximum impact on performance

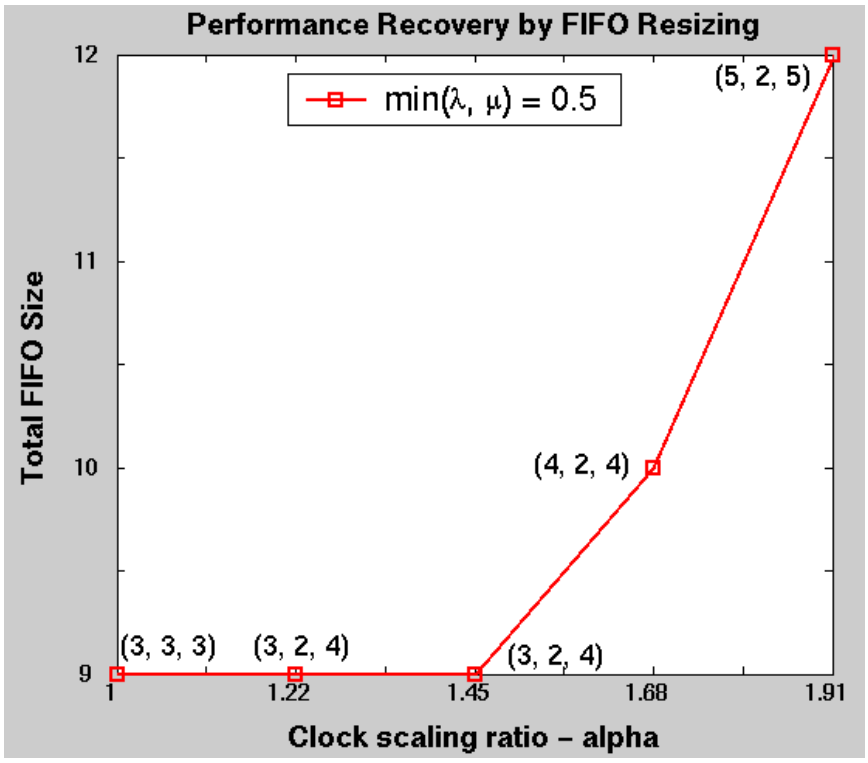


Performance recovery: FIFO resizing

- **3 stage FIFO channel is analyzed**
  - Each stage is of size 3 as a baseline for single clock design
- **The design methodology**
  - FIFOs are voltage and frequency scaled to save power
    - Results in throughput loss
  - FIFOs are resized to recover the performance loss
- **Trade-off between FIFO sizes and clock frequency**

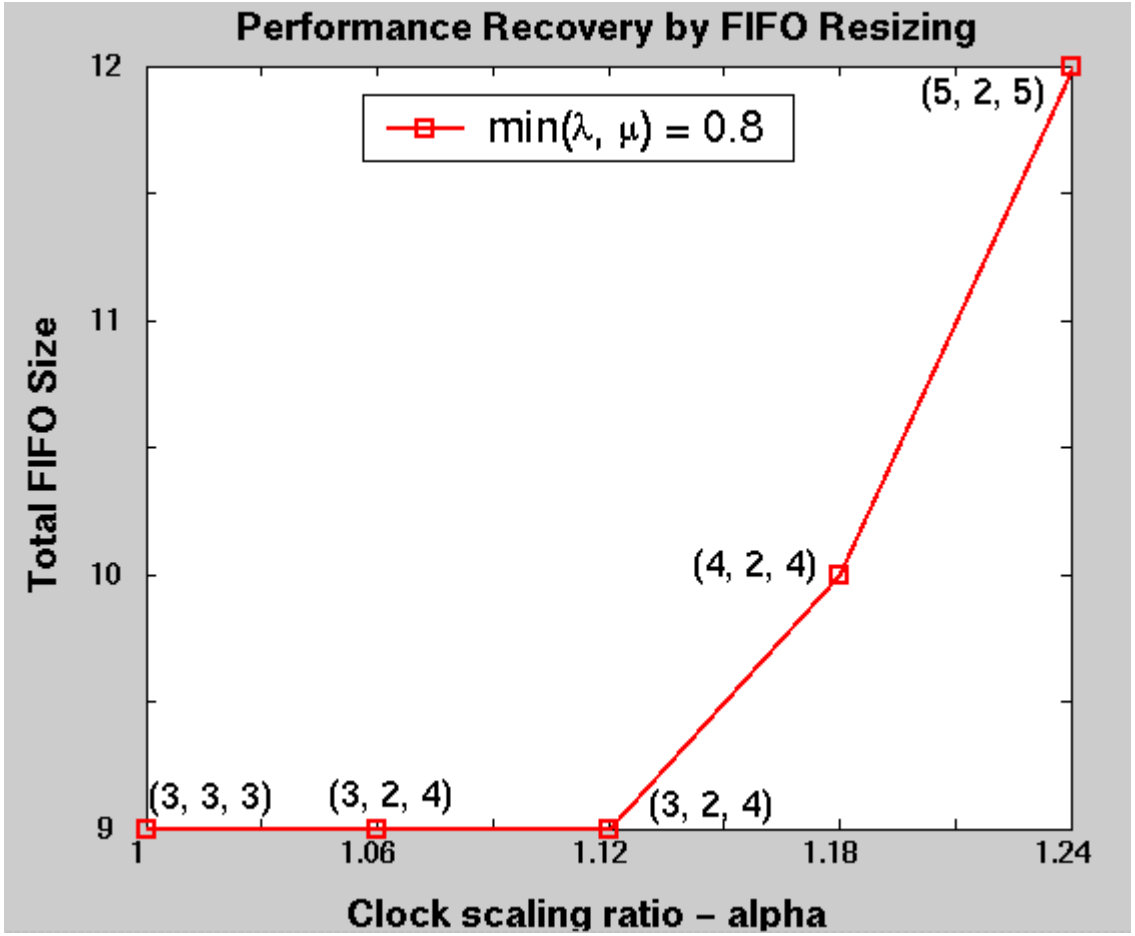
Performance recovery: FIFO resizing





- **Total FIFO sizes increase as a increases**
  - The sizes of beginning and end FIFOs are most important

Performance recovery: FIFO resizing



Power consumption analysis

- **Memory model used for the FIFO**
  - Power consumption in a FIFO is proportional to no. of data words

$$Power \propto \sqrt{n}$$

- **FIFOs have different read and write clocks**

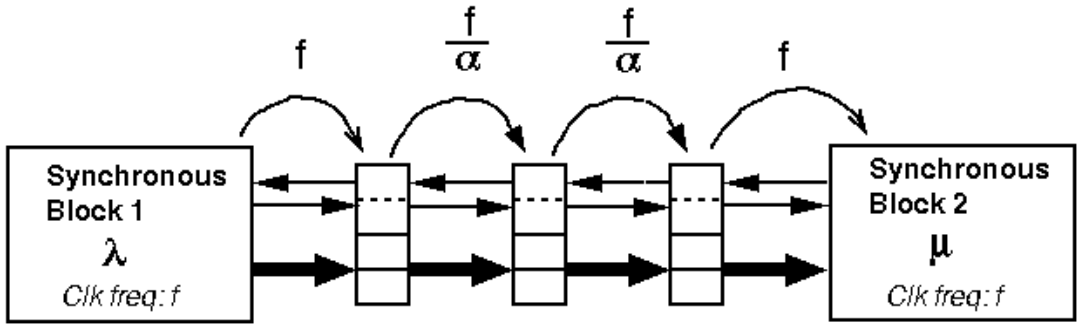
$$Power \propto \sqrt{n} \cdot V^2 \cdot (f_{in} + f_{out})$$

- **For a single clock design**

$$Power_u = K \cdot \sum_{i=1}^{N} \sqrt{n_{iu}} \cdot V^2 \cdot (f + f)$$

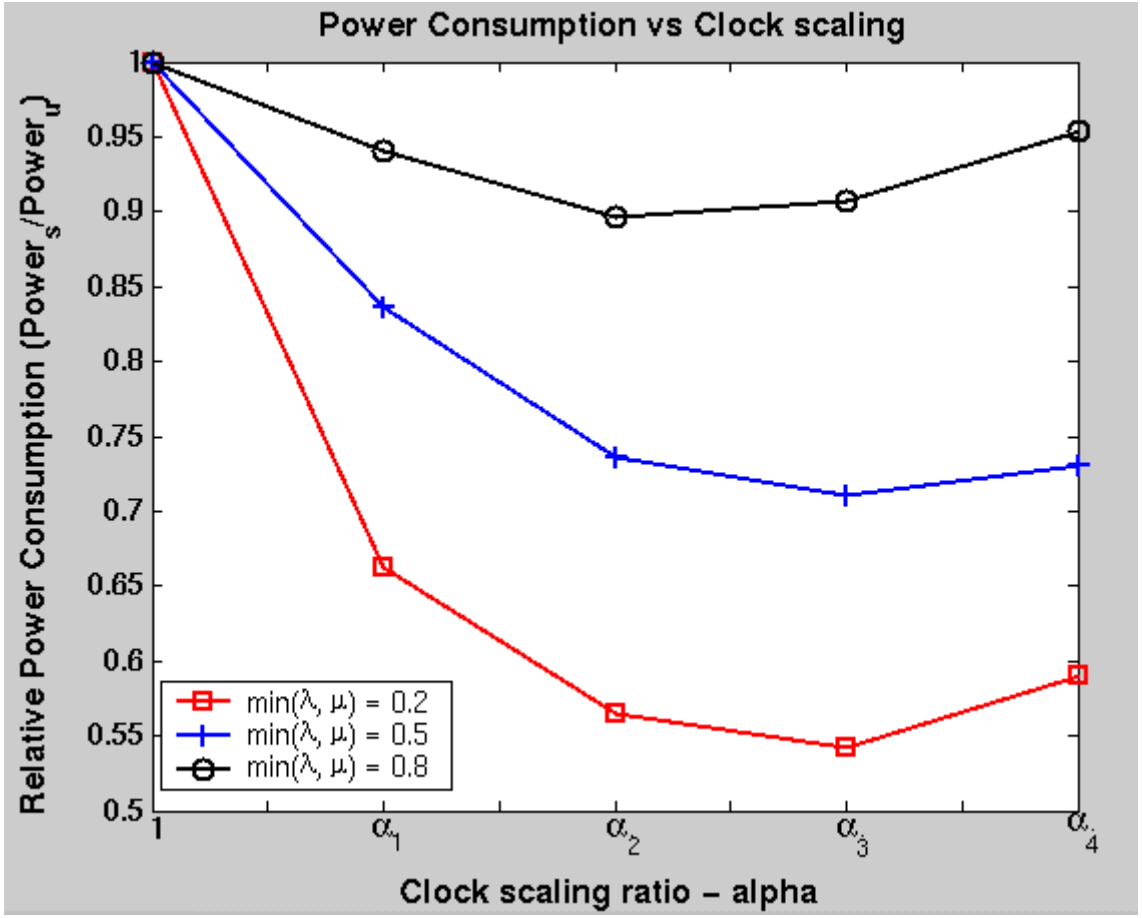
Power consumption analysis

- **For voltage and frequency scaled channel**
  - Voltages for the first and last FIFO are not scaled



$$
\begin{aligned}
Power_s &= K \cdot \sqrt{n_{1s}} \cdot V^2 \cdot \left(f + \frac{f}{\alpha} \right) \\
&+ K \cdot \sum_{i=2}^{N-1} \sqrt{n_{is}} \cdot \left(\frac{V}{\sqrt{\alpha}} \right)^2 \cdot \left(\frac{f}{\alpha} + \frac{f}{\alpha} \right) \\
&+ K \cdot \sqrt{n_{Ns}} \cdot V^2 \cdot \left(\frac{f}{\alpha} + f \right)
\end{aligned}
$$

Power consumption analysis



Power saving decreases as min(l, m) increases!

Conclusions

- **Explored power-performance trade-off**
  - for interconnect channel containing multiple FIFO stages
- **Transfer blocking causes l_{eff} and m_{eff} to be less than l and m respectively**
- **Theoretical upper bound on a given by**

$$\alpha_{max} = \frac{1}{min(\lambda_{eff}, \mu_{eff})}$$

with no performance degradation

  - Practical upper bound on α is less than α_{max} due to transfer blocking

Conclusions

- **Voltage and clock scaling saves power**
  - Performance is lost due to clock scaling
  - Performance is recovered by resizing of the FIFOs in the channel
  - Resizing of the FIFOs at the beginning and end are critical
  - FIFOs in the middle should remain minimum size
- **Max power savings of 45.8%, 28.9%, 11.3%**
  - for min(λ,μ) of 0.2, 0.5 and 0.8 respectively
- **Power saving decreases as min(l, m) increases**