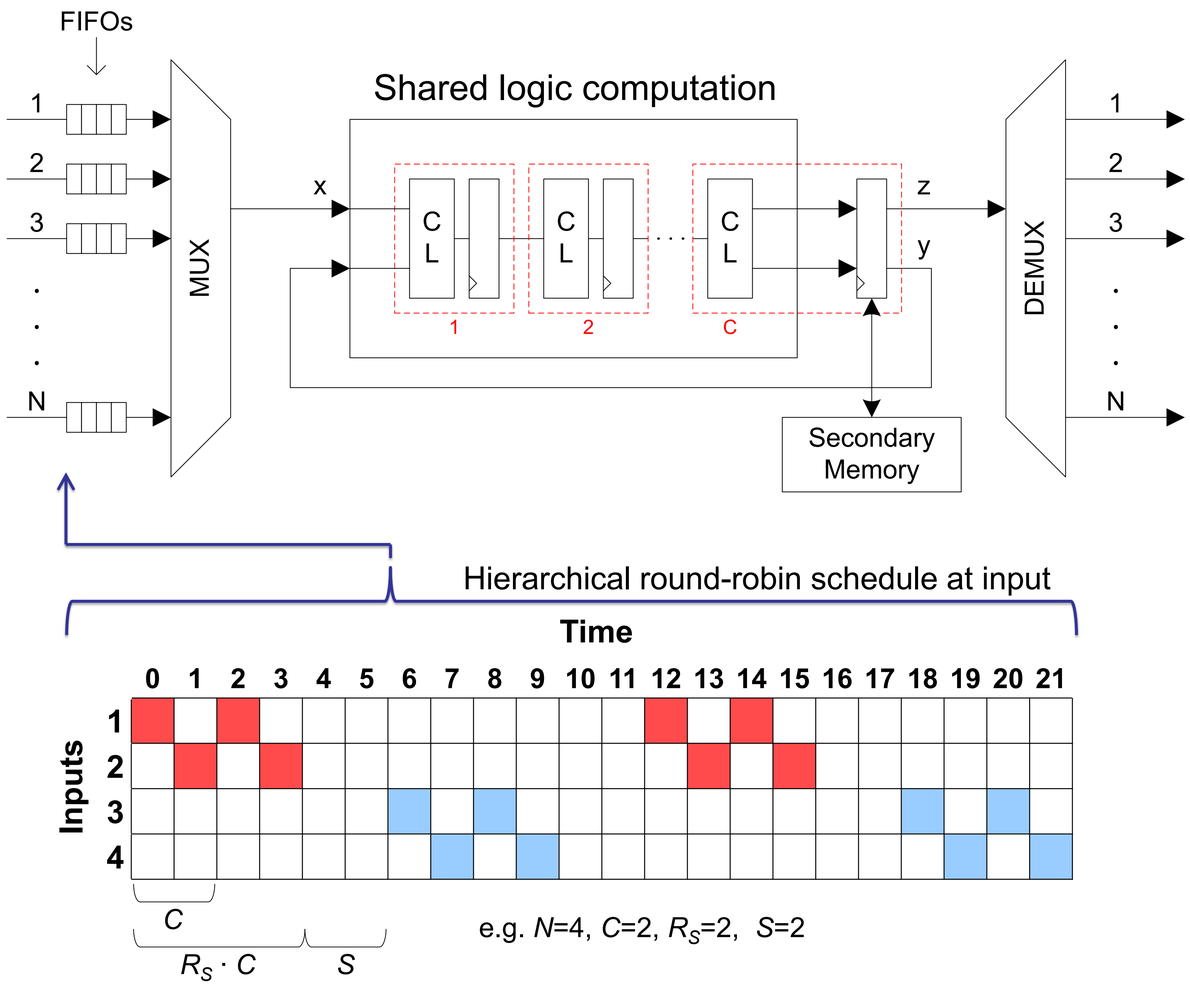
Performance Modeling of Virtualized Custom Logic Computations

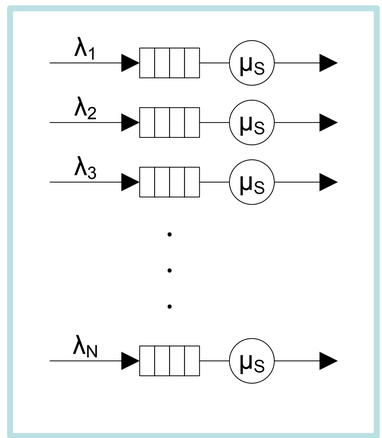
Michael J. Hall, Roger D. Chamberlain
Department of Computer Science and Engineering
Washington University in St. Louis
School of Engineering & Applied Science

Supported by
NSF Grant CNS-0931693 and Exegy, Inc.

Virtualized Logic Computation



Queueing Model



Model definition:

Tput, *Latency*, *Occupancy* = *f* (*Circuit*, *Tech*, *C*, *N*, *S*, R_S, λ)

Variable	Definition
Circuit	Logical circuit description (e.g. written in Verilog or VHDL)
Tech	Target technology (e.g. FPGA or ASIC)
C	Pipeline depth (fine-grain contexts)
N	Total number of contexts (requires secondary memory if *N*>*C*)
S	Cost of a context switch (for secondary memory)
R_S	Scheduling period (number of rounds of *C* contexts that execute before context-switching to secondary memory)
λ	Arrival rate (e.g. data elements per second)

M/D/1 modeling expressions

Service rate of each virtual logic computation:

Total achievable throughput:

$$\mu_s = \frac{R_S}{(R_S N + SN/C)\cdot t_{CLK}} \ \text{elements/s}$$

$$T_{TOT} = N\cdot\mu_s = \frac{R_S}{(R_S + S/C)\cdot t_{CLK}} \ \text{elements/s}$$

New equations in red boxes

Waiting time expressions: (Latency)

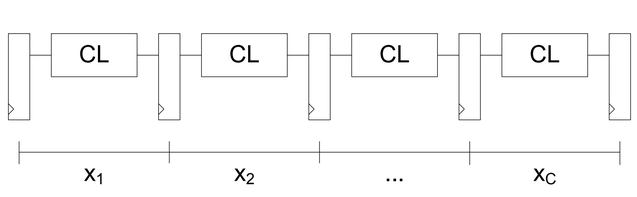
$$W_q = \frac{1}{\mu_s}\cdot\frac{\rho}{2(1-\rho)} \qquad W_s = \frac{1}{\mu_s} = C\cdot t_{CLK}$$

$$W_h = \frac{(R_S(N-C)+SN/C)^2\cdot t_{CLK}}{2(R_S N + SN/C)} \qquad W = W_q + W_h + W_s$$

Queue occupancy:

$$N_q = \lambda\cdot W_q$$

Calibration



Variable	Definition
X	Random variable of stage-to-stage delay for C random samples, x_1, x_2, …, x_C
C	Pipeline depth
t_{CL}	Total comb. logic delay
k	Curve fit parameter

Upper bound on t_{CLK} from order statistics:

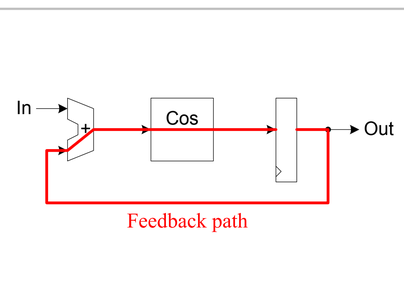
$$t_{CLK} = E[\max(X)] \le \mu_X + \sigma_X\sqrt{C-1}$$

Clock period sub-model:

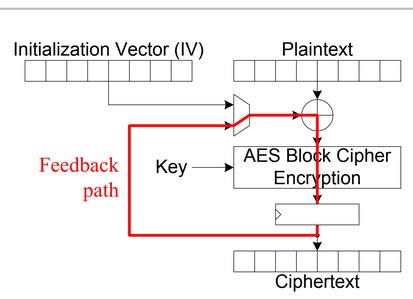
$$t_{CLK} = t_{CL}/C + k\cdot\sqrt{C-1}$$

$$T_{TOT} = \frac{1}{t_{CLK}}$$

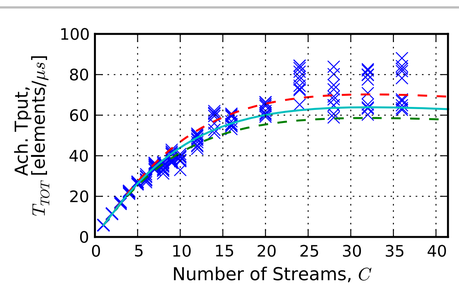
Synthetic Cosine Application with Feedback (20 Taylor-series terms)



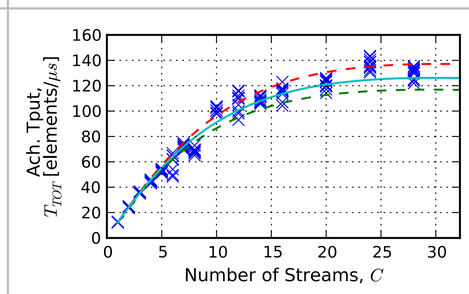
AES Encryption Cipher in CBC Block Mode (14 fully unrolled rounds)



FPGA

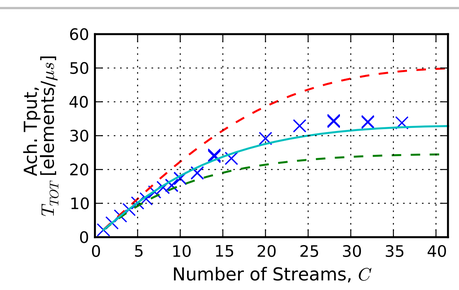


$$t_{CLK} = 8.6\,\tfrac{\text{ns}}{\text{term}}\cdot\frac{N_t}{C} + (1.9\,\text{ns} - 4.7\,\tfrac{\text{ps}}{\text{term}}\cdot N_t)\cdot\sqrt{C-1}$$

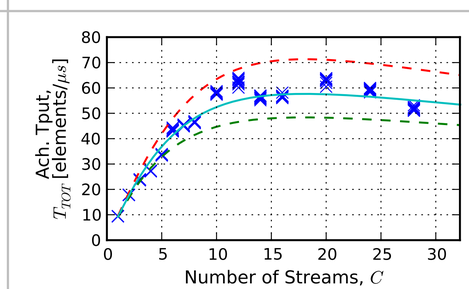


$$t_{CLK} = 5.7\,\tfrac{\text{ns}}{\text{rnd}}\cdot\frac{N_r}{C} + (1.7\,\text{ns} - 55\,\tfrac{\text{ps}}{\text{rnd}}\cdot N_r)\cdot\sqrt{C-1}$$

ASIC

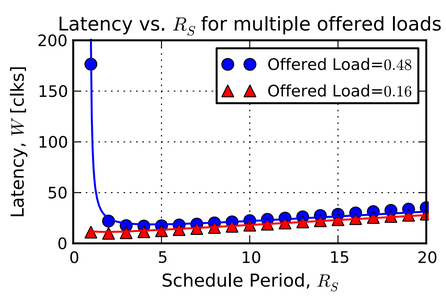


$$t_{CLK} = 2.3\,\tfrac{\text{ns}}{\text{term}}\cdot\frac{N_t}{C} + (3.2\,\text{ns} - 9.8\,\tfrac{\text{ps}}{\text{term}}\cdot N_t)\cdot\sqrt{C-1}$$



$$t_{CLK} = 7.8\,\tfrac{\text{ns}}{\text{rnd}}\cdot\frac{N_r}{C} + (4.4\,\text{ns} - 122\,\tfrac{\text{ps}}{\text{rnd}}\cdot N_r)\cdot\sqrt{C-1}$$

Model Validation



- Offered load is the ratio of aggregate arrival rate to the peak service rate of the system and is equal to $N{\cdot}\lambda{\cdot}t_{CLK}$.
- Points are discrete event simulation
- Curves are analytic expressions

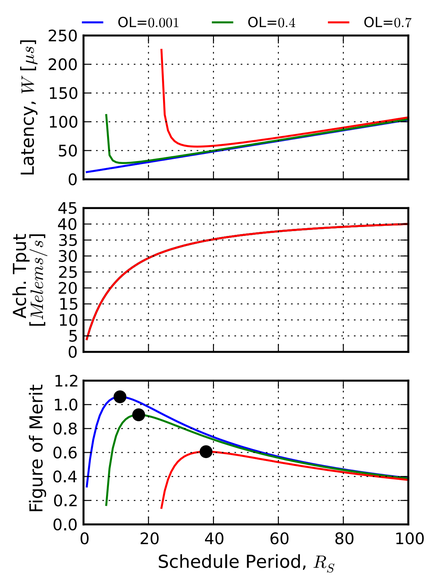
N = 8, *C* = 4, *S* = 4, t_{CLK} = 100 MHz

Ways to Use the Model and Results

Example Design - Case 1

Given:
- *Circuit*=COS, *Tech*=FPGA, *N*=100, *C*=10, *S*=100, *OL* varies

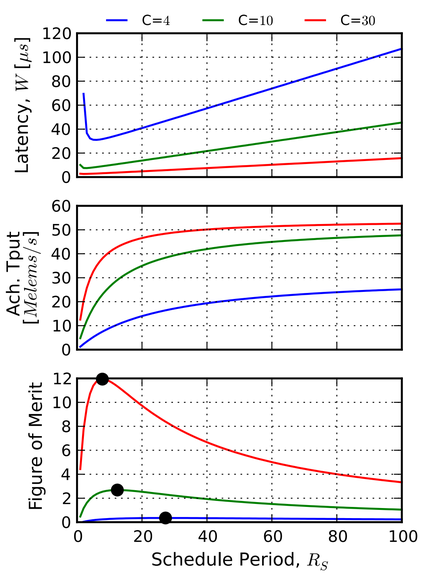
Design Params: R_S
Optimize: *FoM*=*Tput*/*Latency*



Example Design - Case 2

Given:
- *Circuit*=AES, *Tech*=ASIC, *N*=60, *S*=100, *λ*=30 Kelems/s

Design Params: *C*, R_S
Optimize: *FoM*=*Tput*/*Latency*



Optimization of Figure of Merit