Optimal pipeline depth

Anshuman Gupta

Timing Path

anshuman@bb-01:~/research/deploy_hw_hacker (104,47)

anshuman@bb-01:~/researc...

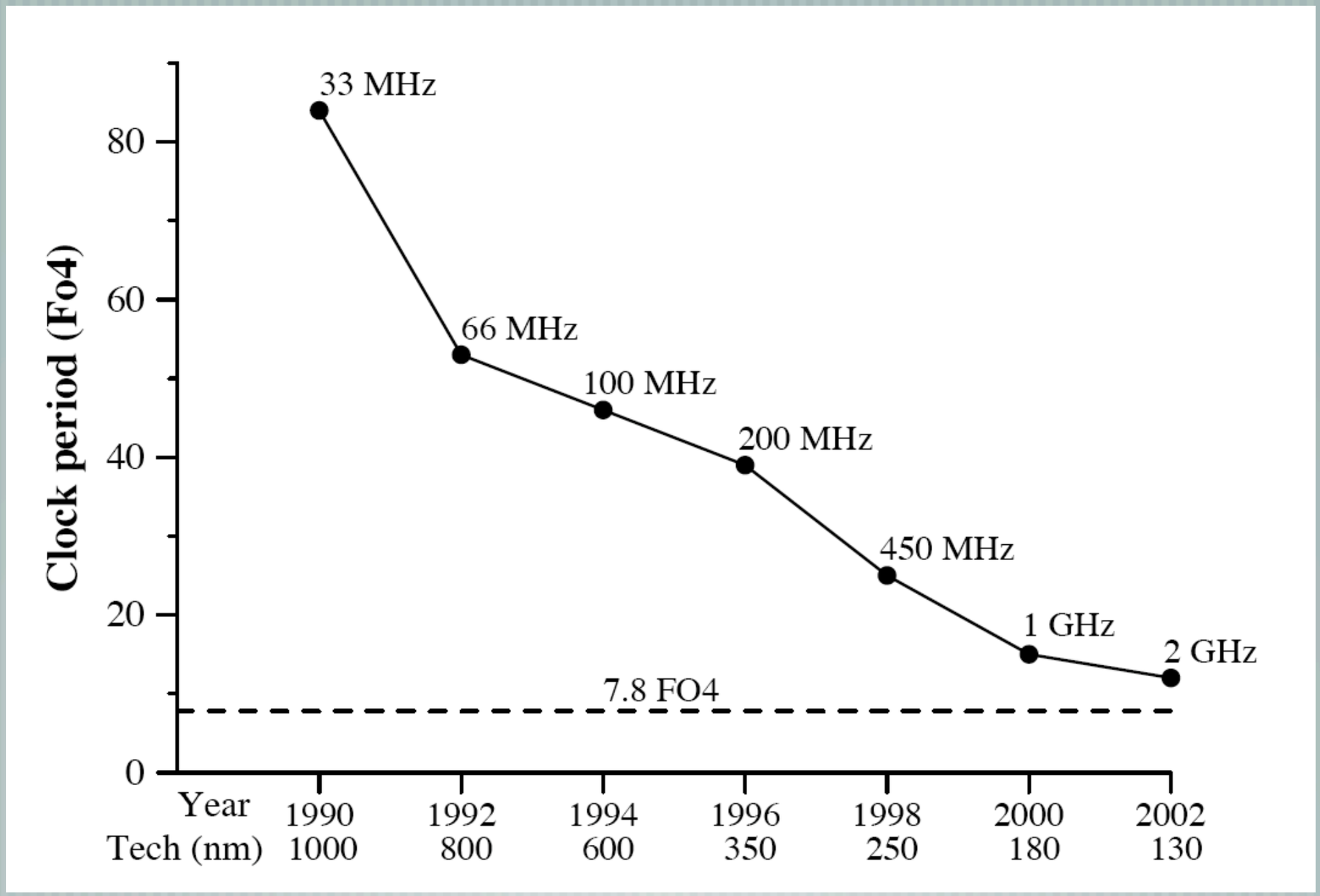
```
Operating Conditions: scx3_tsmc_cln90god_hvt_tt_1p2v_25c   Library: scx3_tsmc_cln90god_hvt_tt_1p2v_25c
Wire Load Model Mode: top

  Startpoint: router/need_out_1_3_reg
              (rising edge-triggered flip-flop clocked by my_clock)
  Endpoint: router/fifo_3_0/data/el0/el_out_r2
            (rising edge-triggered flip-flop clocked by my_clock)
  Path Group: my_clock
  Path Type: max

  Point                                                    Incr       Path
  -----------------------------------------------------------------------------
  clock my_clock (rise edge)                             0.0000     0.0000
  clock network delay (ideal)                            0.0000     0.0000
  router/need_out_1_3_reg/CK (DFFHQX8TH)                 0.0000 #   0.0000 r
  router/need_out_1_3_reg/Q (DFFHQX8TH)                  0.0806     0.0806 f
  router/U15/Y (CLKINVX12TH)                             0.0185     0.0992 r
  router/U17/Y (OAI222X4TH)                              0.0602     0.1593 f
  router/U19/Y (NOR4X6TH)                                0.1067     0.2660 r
  router/U16/Y (NAND4X6TH)                               0.0704     0.3364 f
  router/U14/Y (CLKINVX20TH)                             0.0258     0.3623 r
  router/U13/Y (NAND4X8TH)                               0.0437     0.4060 f
  router/U12/Y (NOR4X6TH)                                0.0611     0.4671 r
  router/U11/Y (AND4X8TH)                                0.1082     0.5752 r
  router/fifo_3_0/deque (g_fifo_4_6)                     0.0000     0.5752 r
  router/fifo_3_0/data/deque (g_fifo_data_6)             0.0000     0.5752 r
  router/fifo_3_0/data/el0/deque (g_fifo_elmt_head_6)    0.0000     0.5752 r
  router/fifo_3_0/data/el0/U68/Y (INVX3TH)               0.0393     0.6146 f
  router/fifo_3_0/data/el0/U65/Y (NOR2X8TH)              0.0712     0.6858 r
  router/fifo_3_0/data/el0/U72/Y (CLKBUFX40TH)           0.0632     0.7490 r
  router/fifo_3_0/data/el0/U40/Y (AOI22X4TH)             0.0300     0.7789 f
  router/fifo_3_0/data/el0/U39/Y (OAI2BB1X4TH)           0.0225     0.8014 r
  router/fifo_3_0/data/el0/el_out_r2/D (DFFHQX8TH)       0.0000     0.8014 r
  data arrival time                                                 0.8014

  clock my_clock (rise edge)                             0.8500     0.8500
  clock network delay (ideal)                            0.0000     0.8500
  router/fifo_3_0/data/el0/el_out_r2/CK (DFFHQX8TH)      0.0000     0.8500 r
  library setup time                                    -0.0457     0.8043
  data required time                                                0.8043
  -----------------------------------------------------------------------------
  data required time                                                0.8043
  data arrival time                                                -0.8014
  -----------------------------------------------------------------------------
  slack (MET)                                                       0.0029
                                                                              16,0-1        1%
```

Clock Period trend (FO4)



Experiment Setup

- Calculate the useful work in each stage (FO4)
- Assume naive pipelining for each stage
- Calculate cycle time using -

$$\phi = \phi_{logic} + \phi_{latch} + \phi_{skew} + \phi_{jitter}$$

Symbol	Definition	Overhead
ϕ_{latch}	Latch Overhead	1.0 FO4
ϕ_{skew}	Skew Overhead	0.3 FO4
ϕ_{jitter}	Jitter Overhead	0.5 FO4
$\phi_{overhead}$	Total	1.8 FO4

- Assume $\phi_{overhead}$ remains the same over generations

Optimal logic depth ...

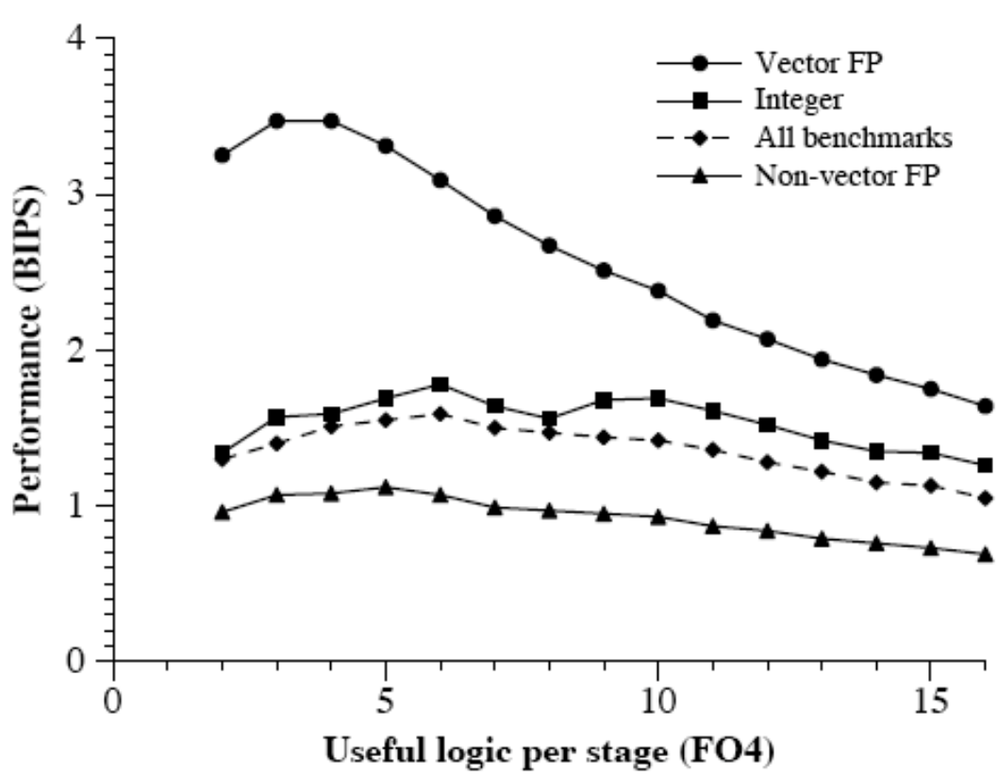
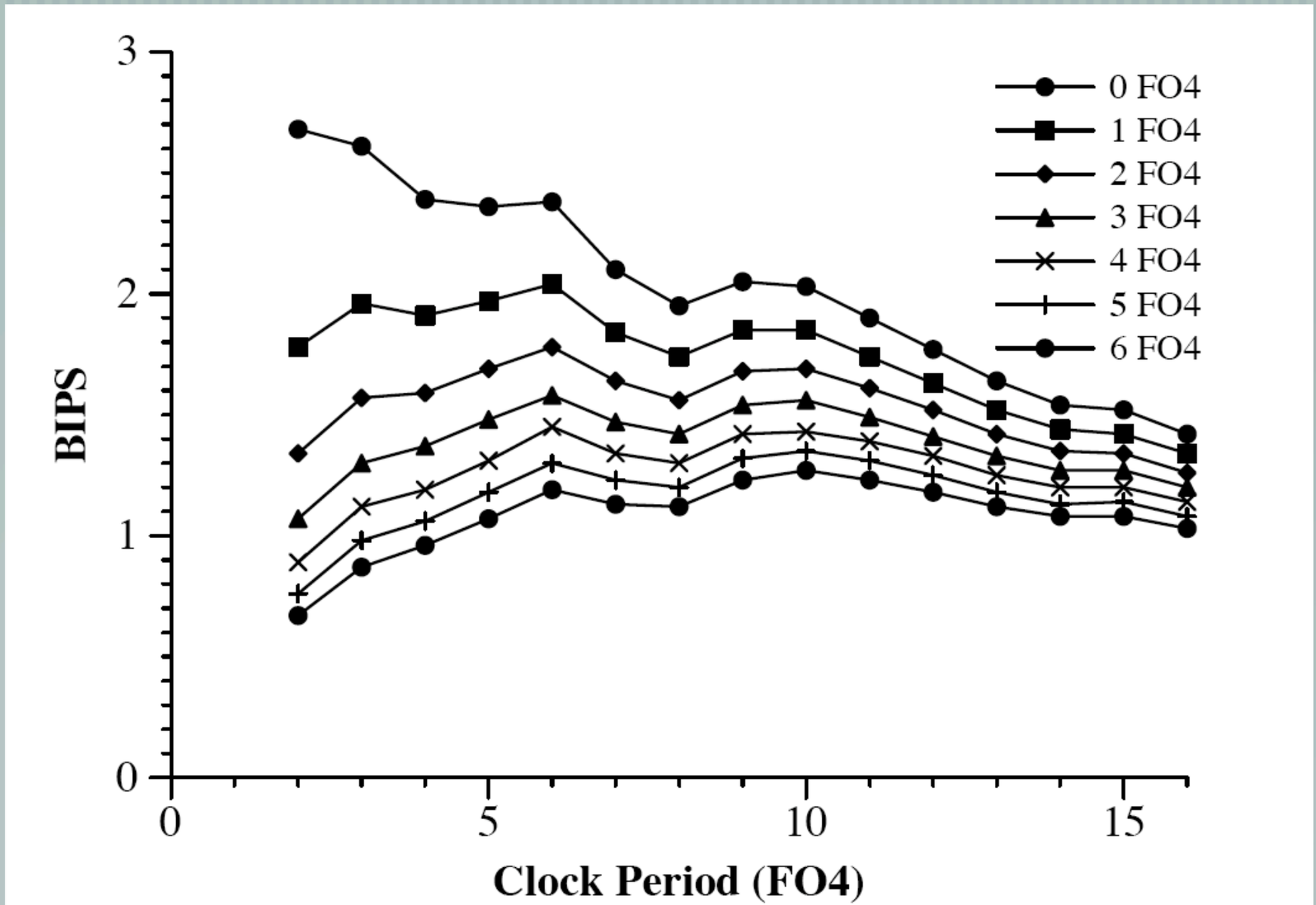
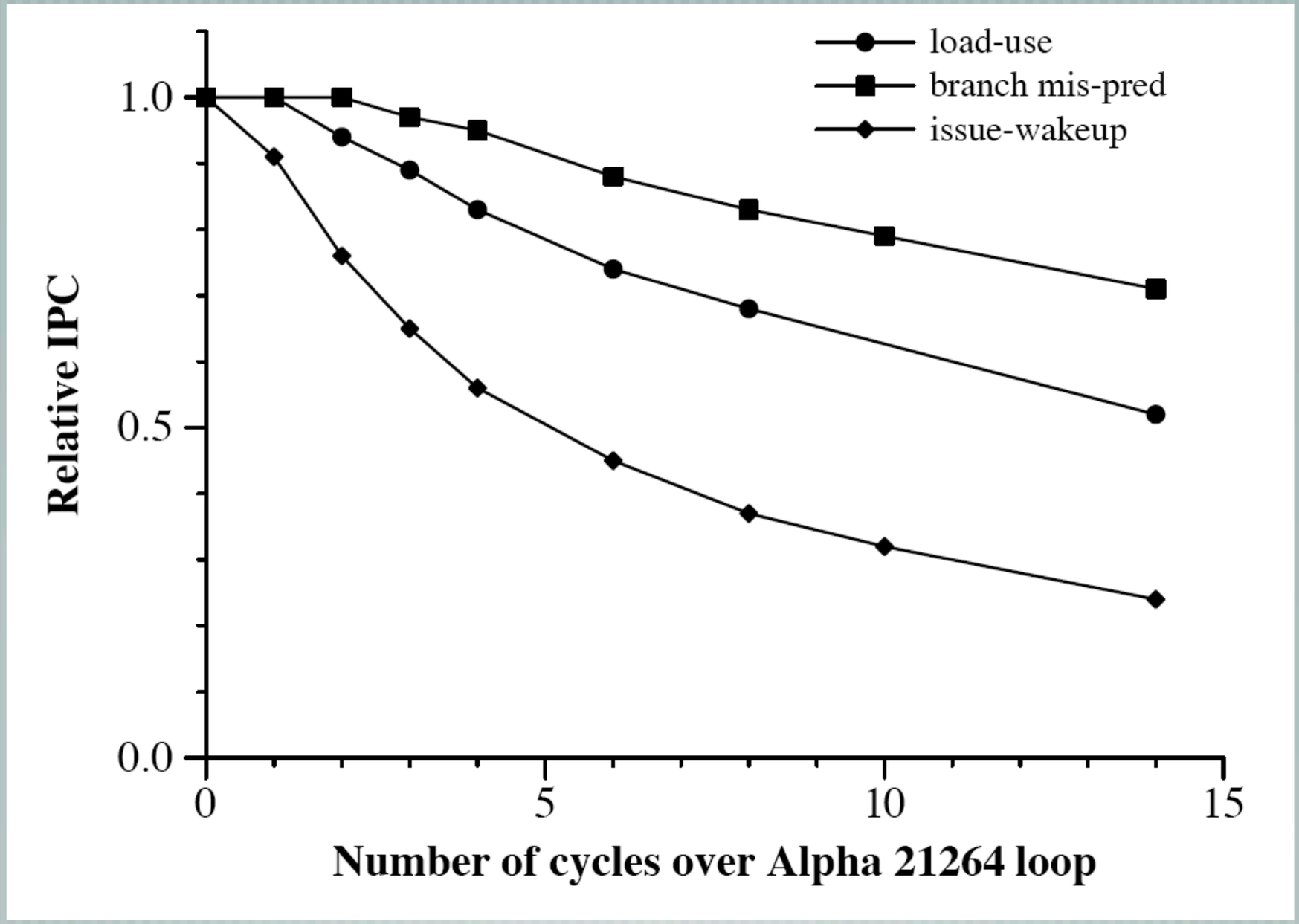


Figure 5: The harmonic mean of the performance of integer and floating point benchmarks, executing on an out-of-order pipeline, accounting for latch overhead, clock skew and jitter. For integer benchmarks best performance is obtained with 6 FO4 of useful logic per stage (ϕ_{logic}). For vector and non-vector floating-point benchmarks the optimal ϕ_{logic} is 4 FO4 and 5 FO4 respectively.

Sensitivity to $\Phi_{overload}$



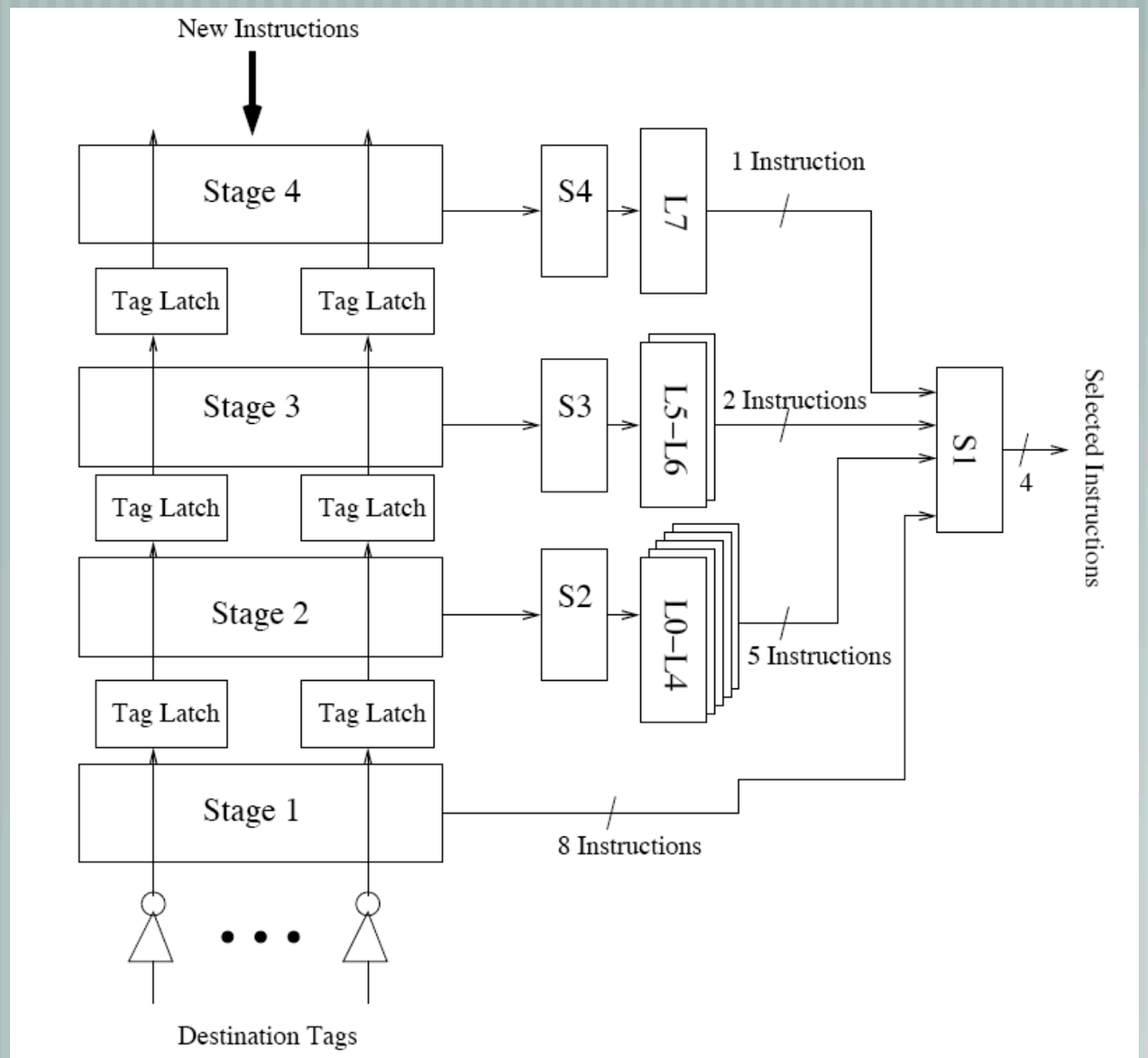
Sensitivity to micro-architectural loops

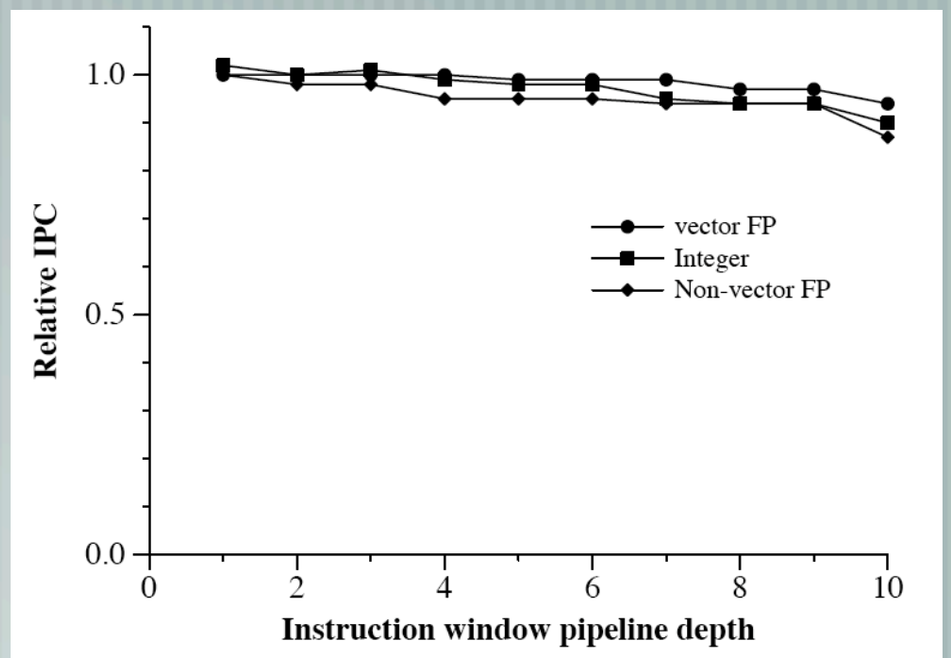


Reclaiming lost performance

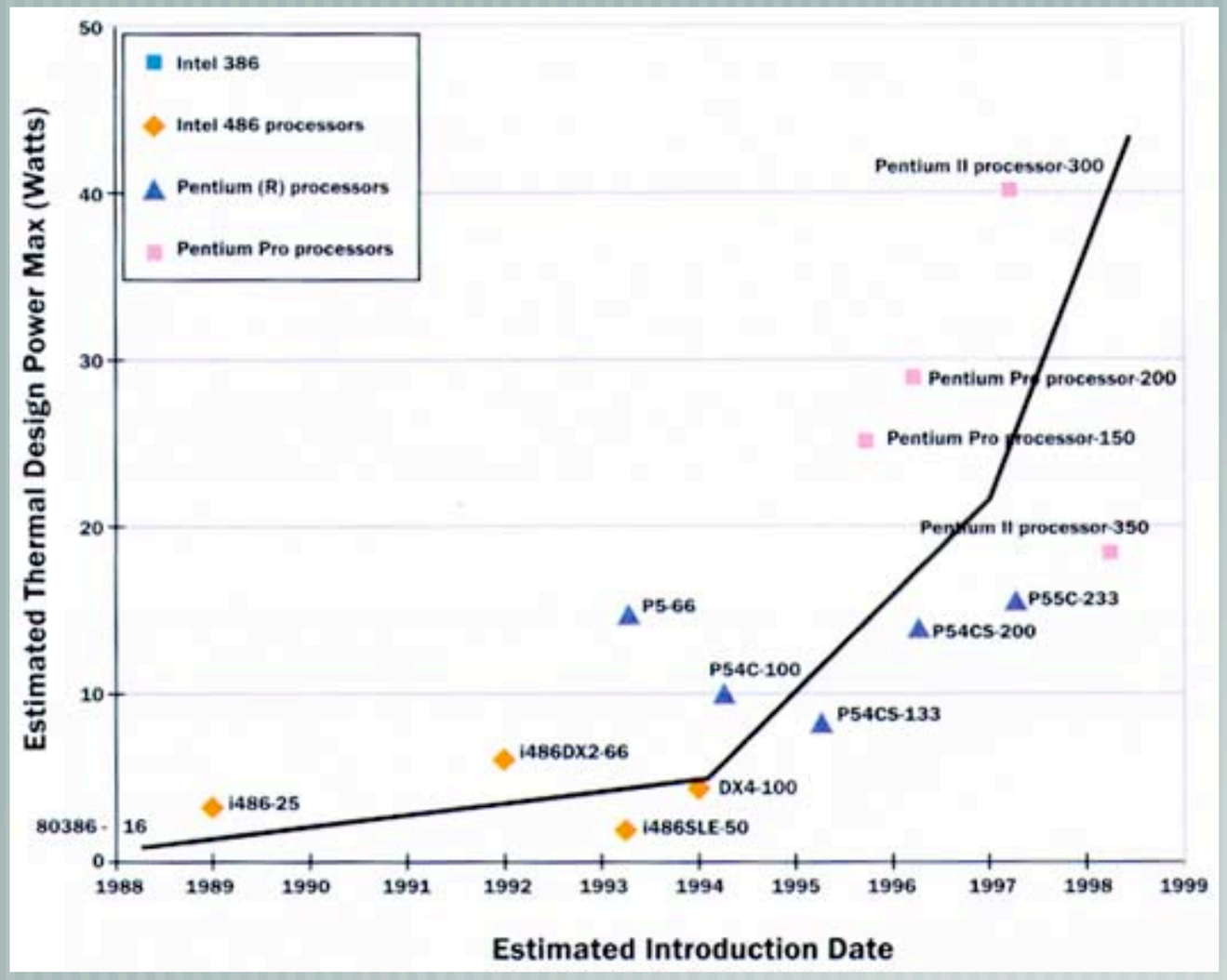
- Architectural loops cause heavy IPC hits with increase in pipeline stages
- Reduce the impact of increased latency to reduce IPC losses
- Use locality, temporality and criticality

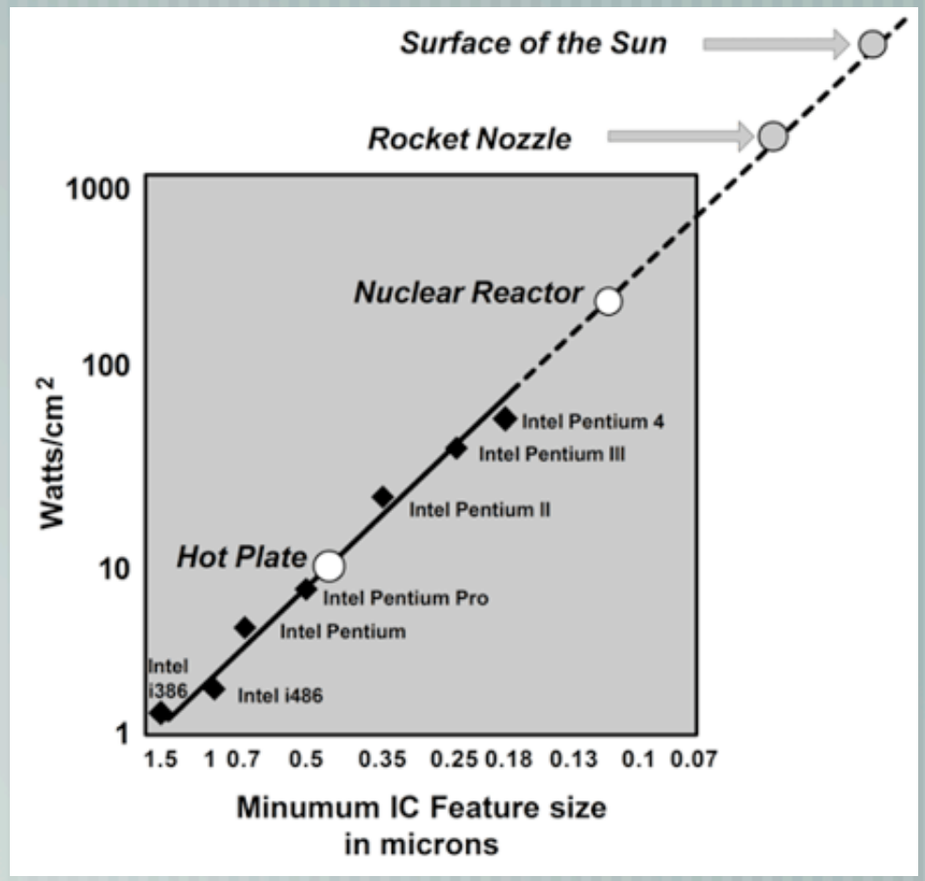
Segmented Instruction Window





Power kicks in





Optimal Power/Performance Pipeline Depth

Performance Equation

$$T/N_I = (\frac{t_o}{\alpha} + \frac{\gamma N_H}{N_I} t_p) + \frac{t_p}{\alpha p} + \frac{\gamma N_H t_o}{N_I} p$$

- Time taken is proportional to -
  - N_H - number of hazards
  - γ - average stalling ratio due to hazards
  - t_0 - overhead delay
  - t_p - total useful pipeline time
- Time taken goes down with -
  - α - degree of superscalar execution
- The relation to number of pipeline stages is not linear

Optimal Power/Performance Pipeline Depth

Power Equation

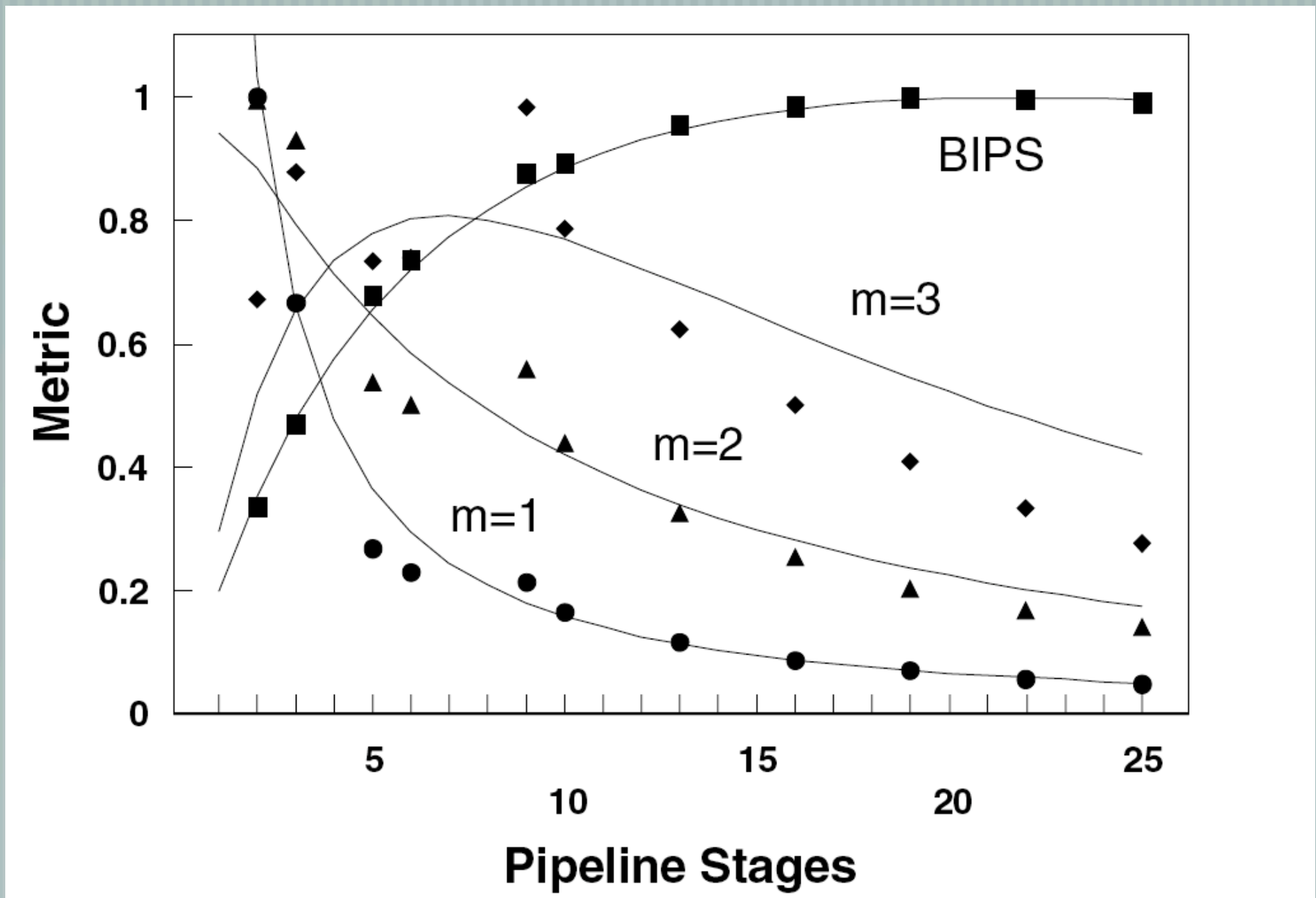
$$P_T = (f_{cg} f_s P_d + P_l) N_L p^{\eta}$$

- Power consumption grows with -
  - P_d - dynamic power, based on
    - f_{cg} - clock gating degree
    - f_s - frequency
  - P_l - leakage power
  - N_L - number of latches
  - p - number of pipeline stages
    - the growth with number of pipeline stages is super-linear due to growth in number of latches by factor η

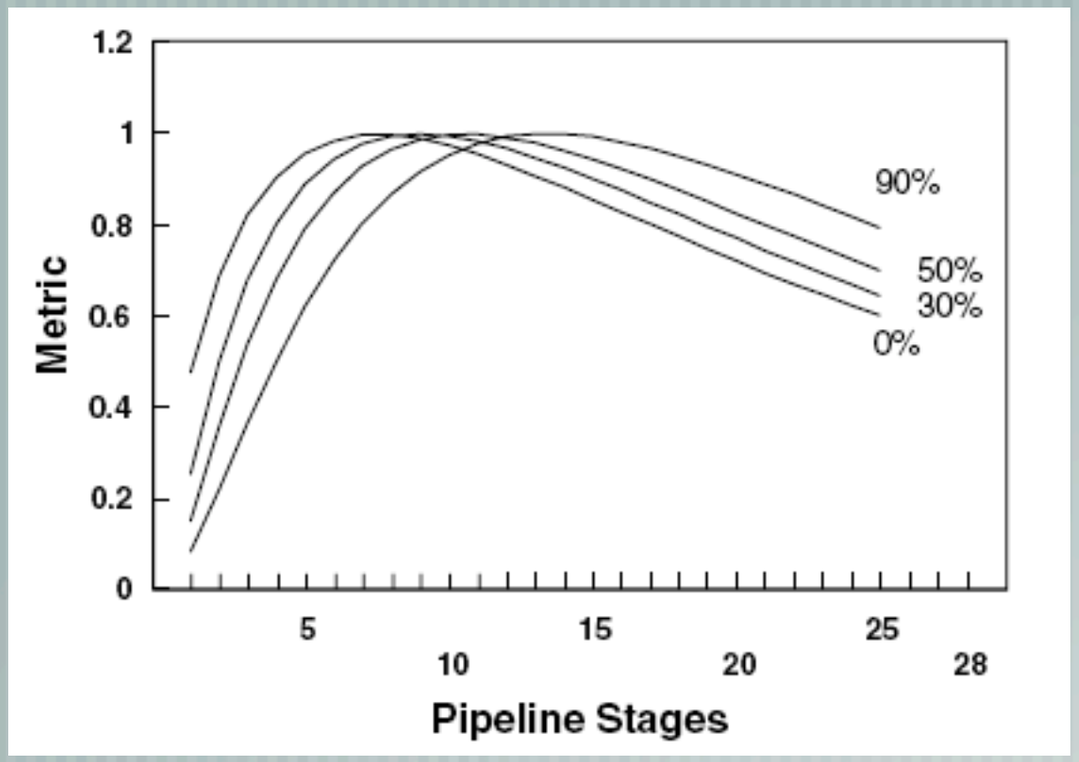
Theoretical results

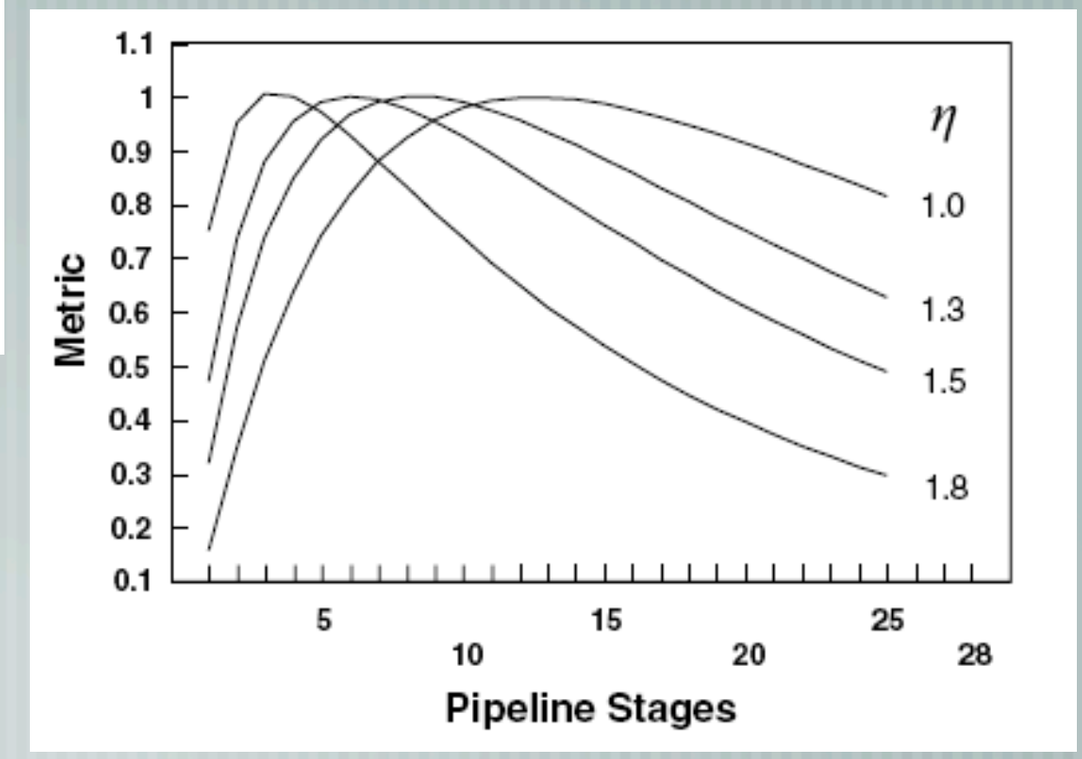
- p_{opt} - optimal number of pipeline stages
- $p_{opt} \propto N_H$
- $p_{opt} \propto \gamma$
- $p_{opt} \propto \alpha$
- $p_{opt} \propto t_p/t_0$

Simulation results



Simulation results



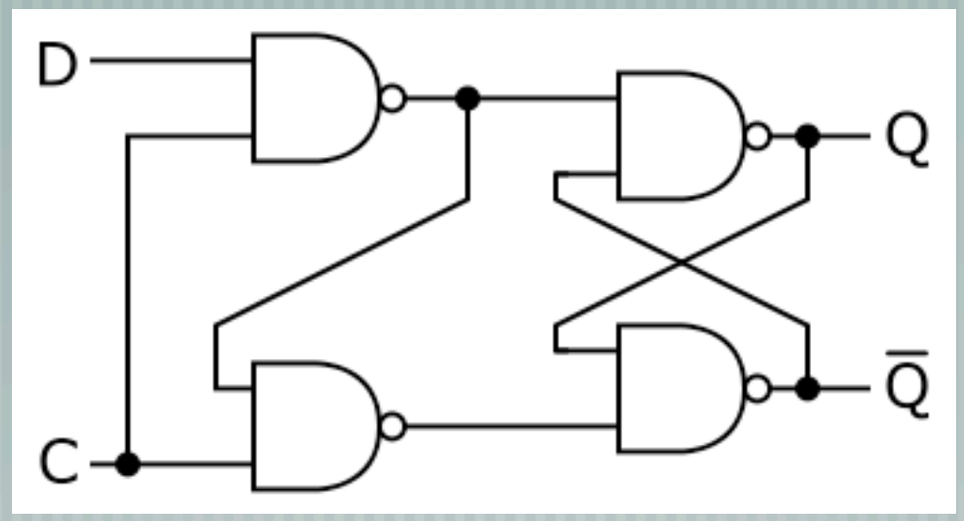


Conclusions

- Power matters!!!
- These results hold true in future only if no major technological breakthrough achieved for changing the equations
- Hazards (N_H) can't be reduced but their impact (γ) can be reduced, traditionally a major research area
- Superscalar architectures prefer shorter pipelines
- Like NUCA we are soon going to see varying access times to L1, register files etc.

APPENDIX slides

Latches and Flip-Flops



- A flip flop is created by using 2 latches (master-slave pair)

Cell Name	Gate Count	Width (um)	Leakage(nW)		
			Min.	Ave.	Max.
DFQD1	6	5.0400	186.0590	204.0590	220.2050

Cell Name	Path	Parameter	Group1	Group2	Group3
			(< 0.0043)pf	(0.0043–0.0799)pf	(> 0.0799)pf
DFQD1	CP to Q	t_{PLH}	0.0667+2.2000*Cload	0.0679+2.0099*Cload	0.0685+2.0037*Cload
		t_{PHL}	0.0699+1.7600*Cload	0.0741+1.1749*Cload	0.0767+1.1128*Cload