Accelerating DFM EDP using the Cell BE microprocessor architecture

F.M. Schellenberg, T. Kingsley, N. Cobb
D. Dudau, R. Chalisani
MENTOR GRAPHICS
www.mentor.com

J. McKibbin, S. McPherson
MERCURY COMPUTING
www.mc.com

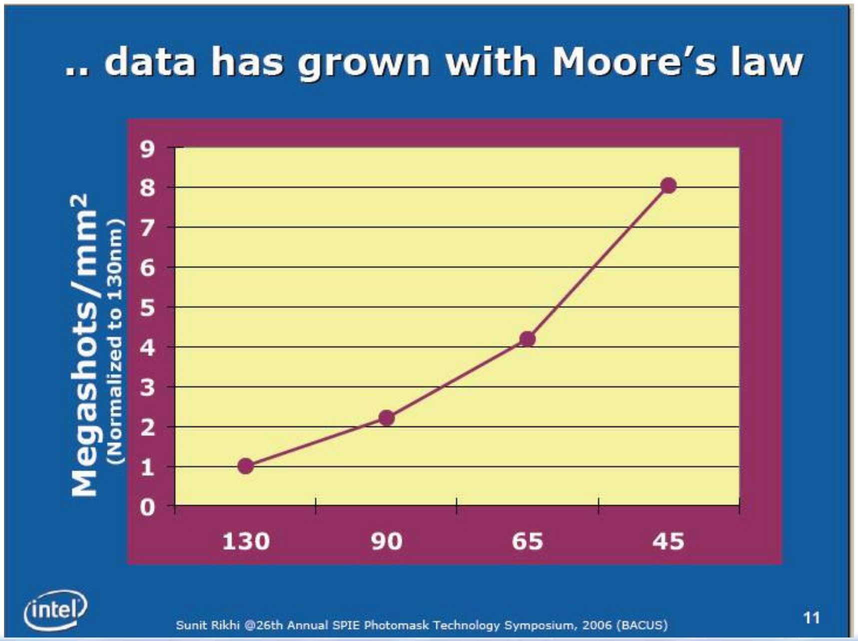
Mentor Graphics®

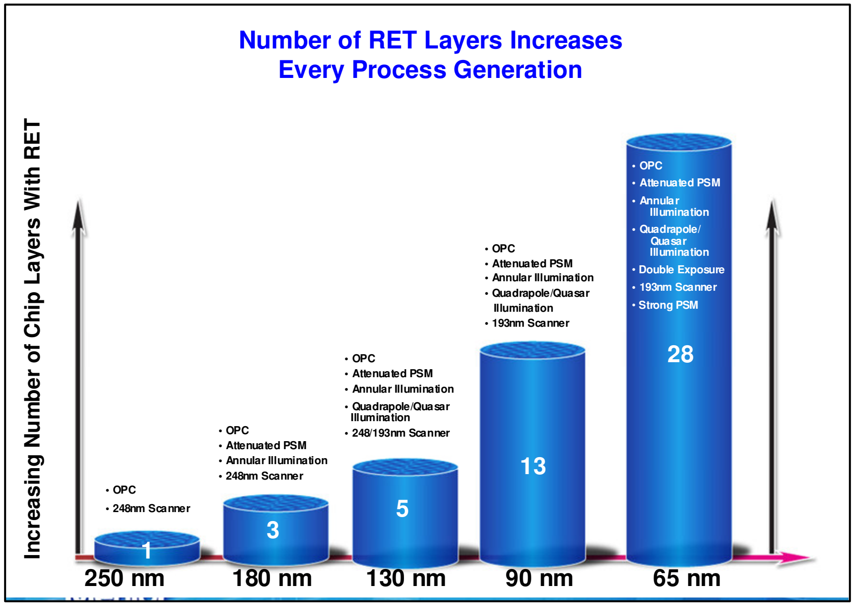
Outline

- Introduction
- Distributed Processing Solutions
- Cell Be platform
  - FFT on the Cell
- Calculating Speedup
- Future applications

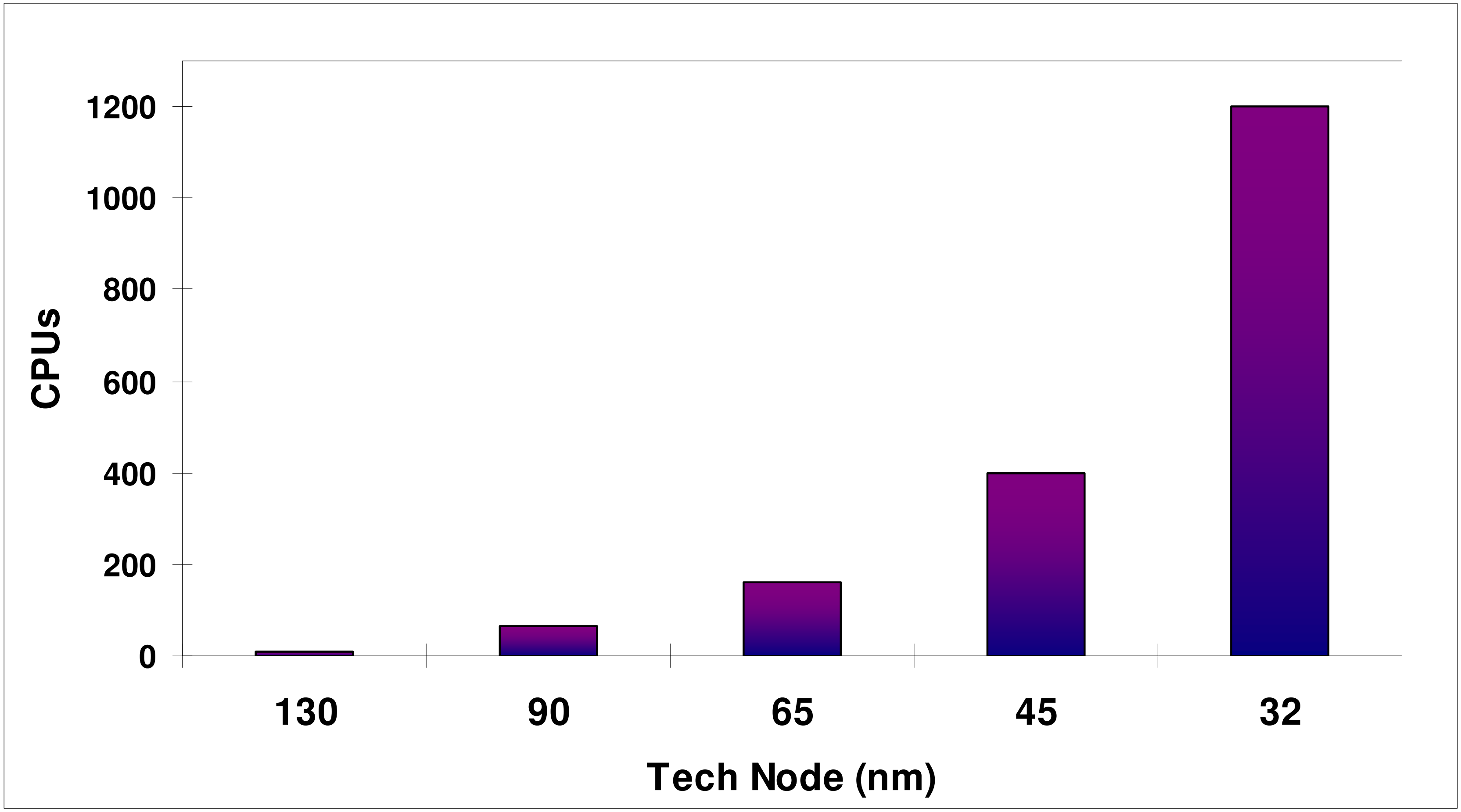
Data size is growing

- Polygon count and O.D. increasing geometrically
- More complex models – more optical kernels, more complex etching, resist, 3D Mask effects. Etc.
- Moving to all-layer RET
- More PW points to correct and verify against

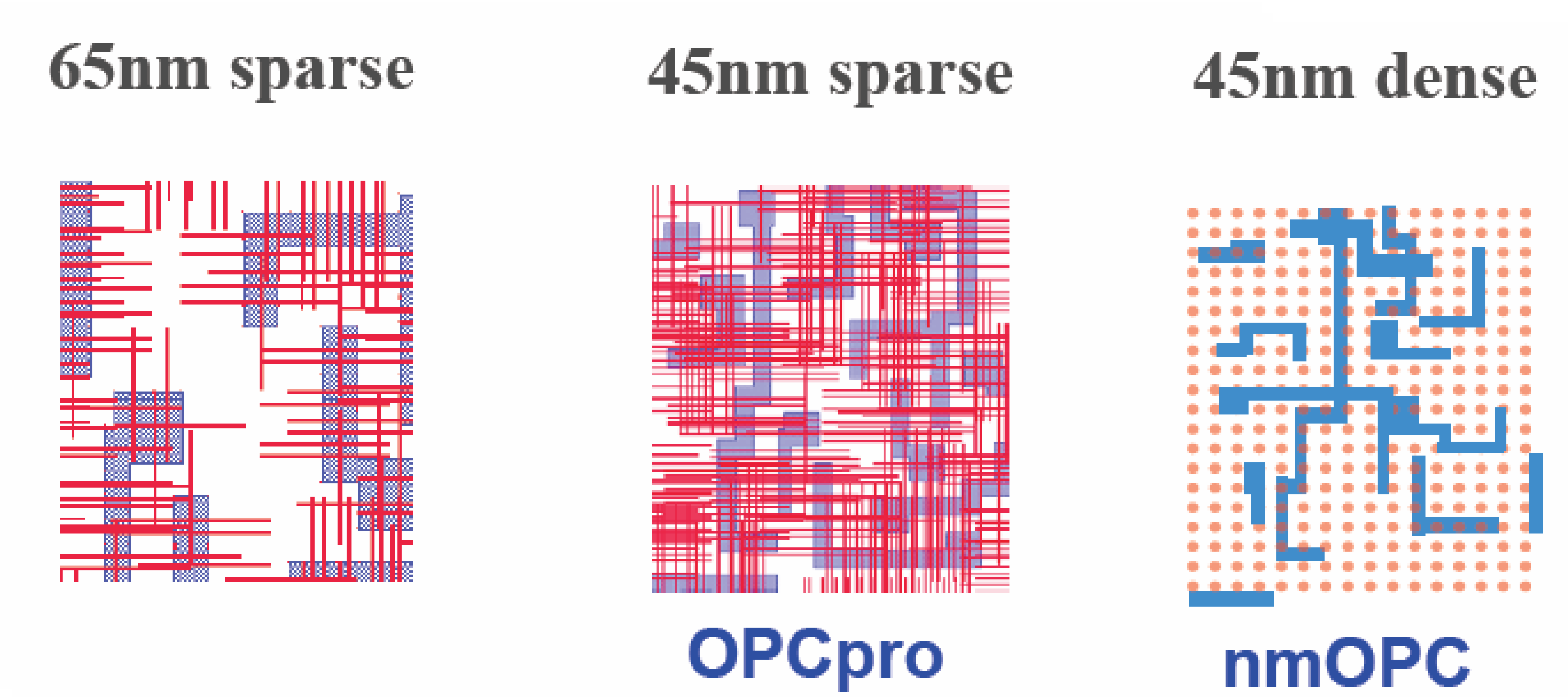




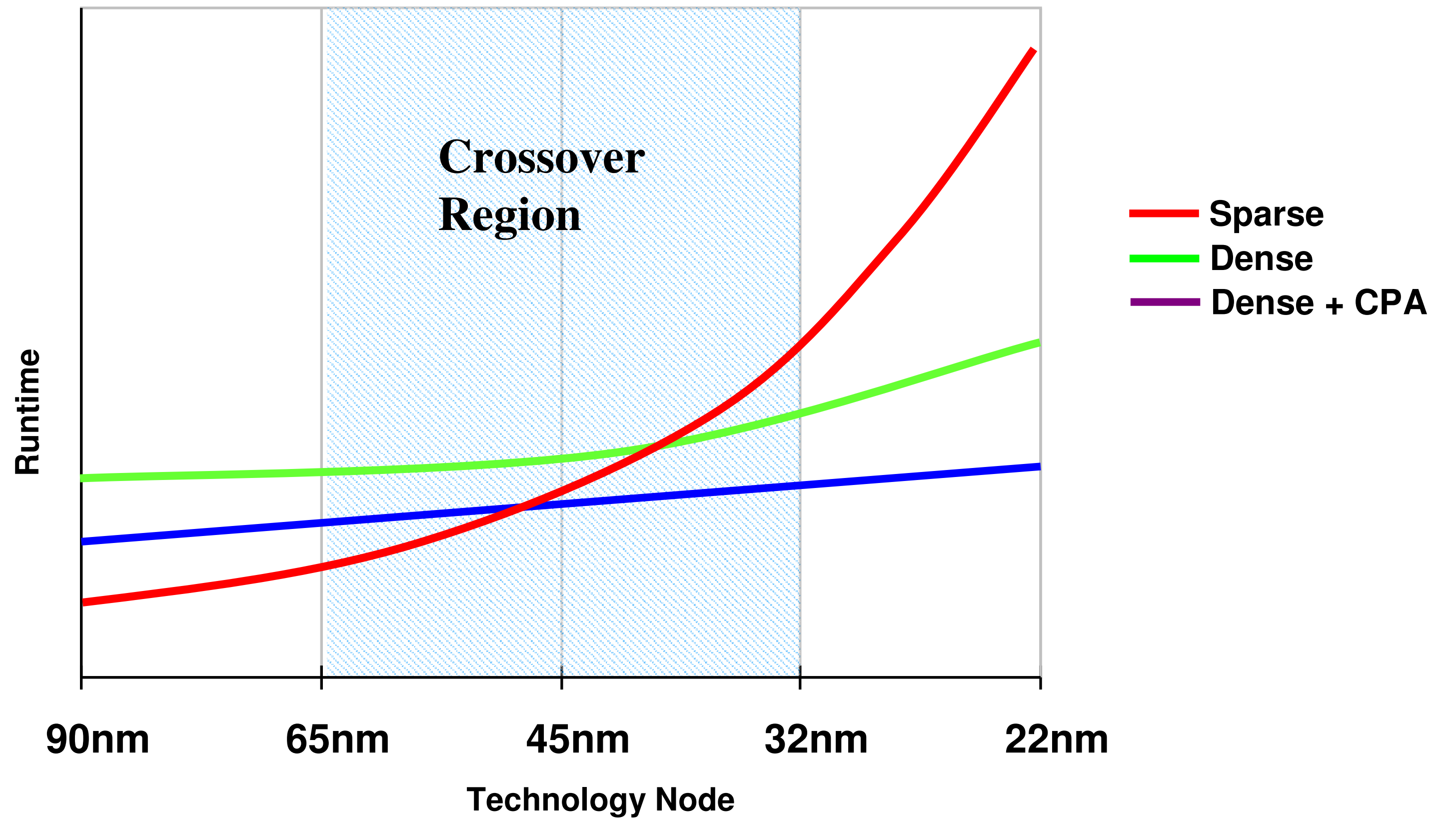
Computation for RET is growing



Computation for RET is growing



OPC Technology Runtimes



Note: Worst case run time on critical layer.

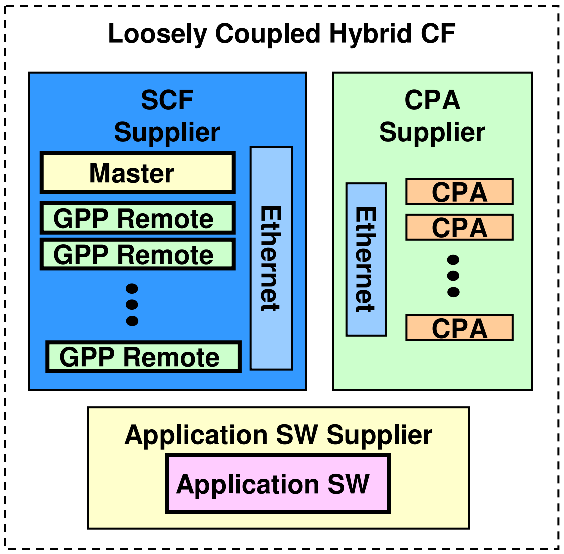
Outline

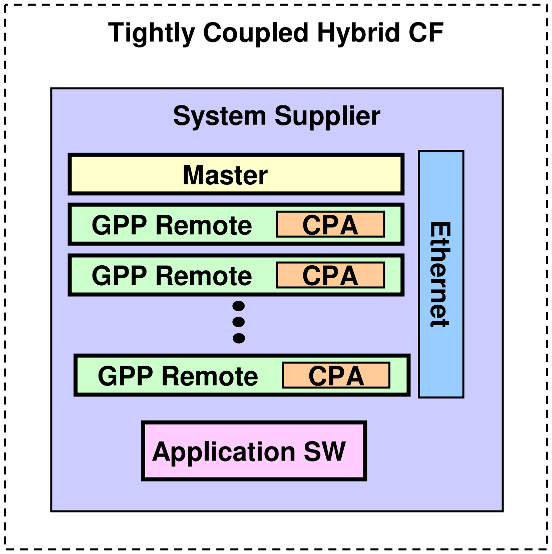
- Introduction
- Distributed Processing Solutions
- Cell Be platform
  - FFT on the Cell
- Calculating Speedup
- Future applications

Two Approaches for Hybrid Compute Farm

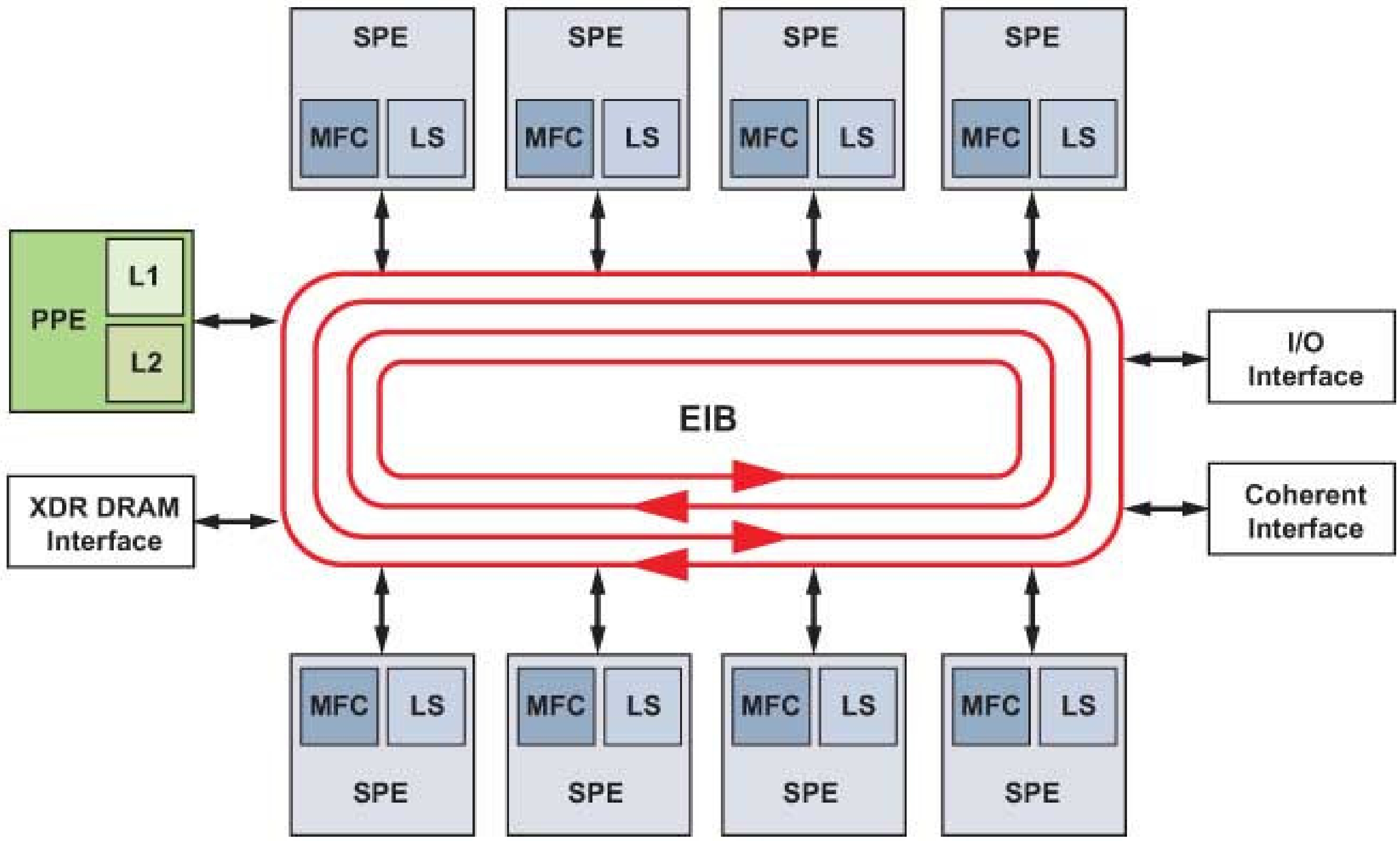
- GPP & CPA physically distinct
  - Can be individually optimized
- GPP↔CPA bandwidth lower

- GPP & CPA physically inseparable
- GPP↔CPA bandwidth high

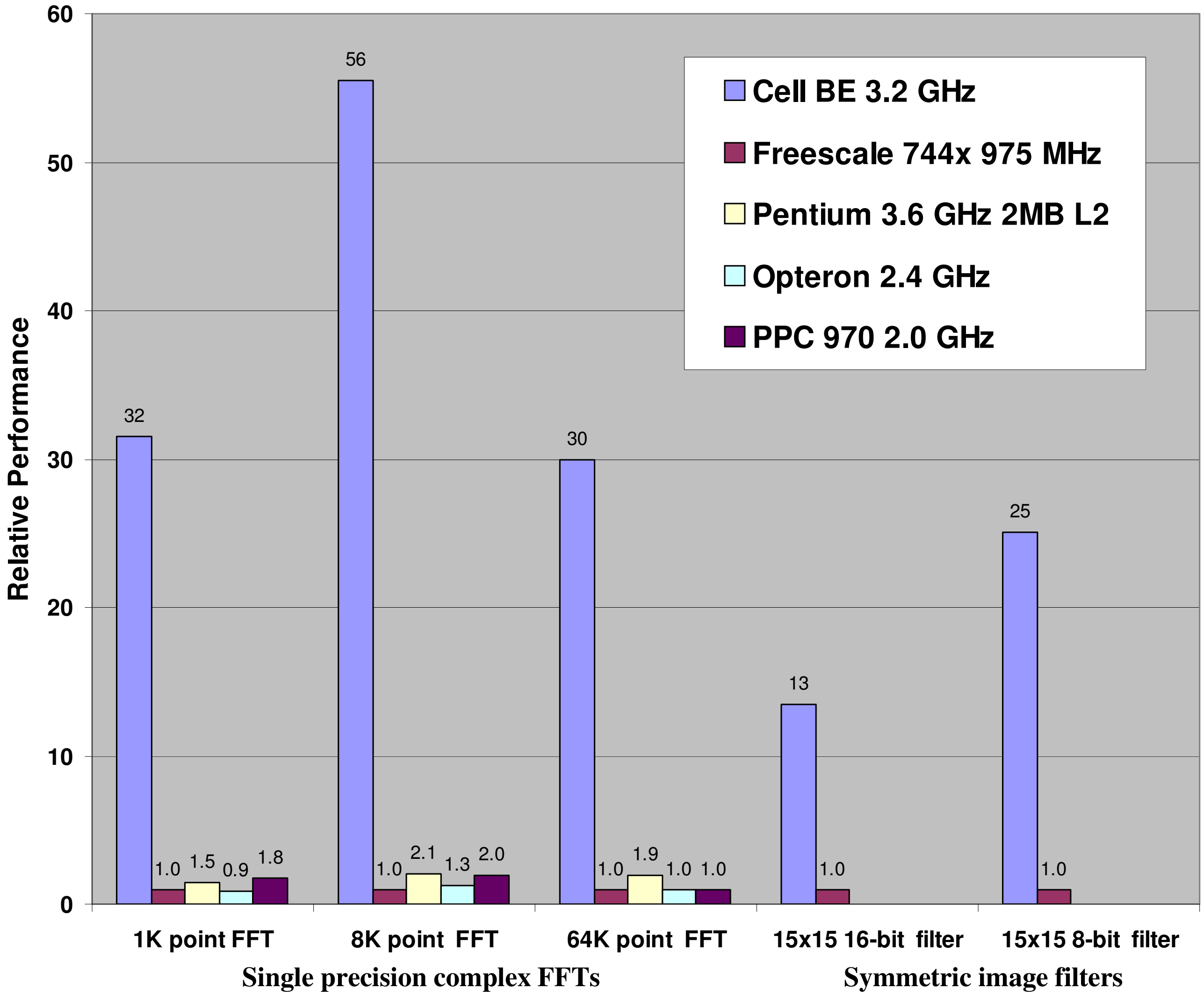




Cell BE Processor Architecture



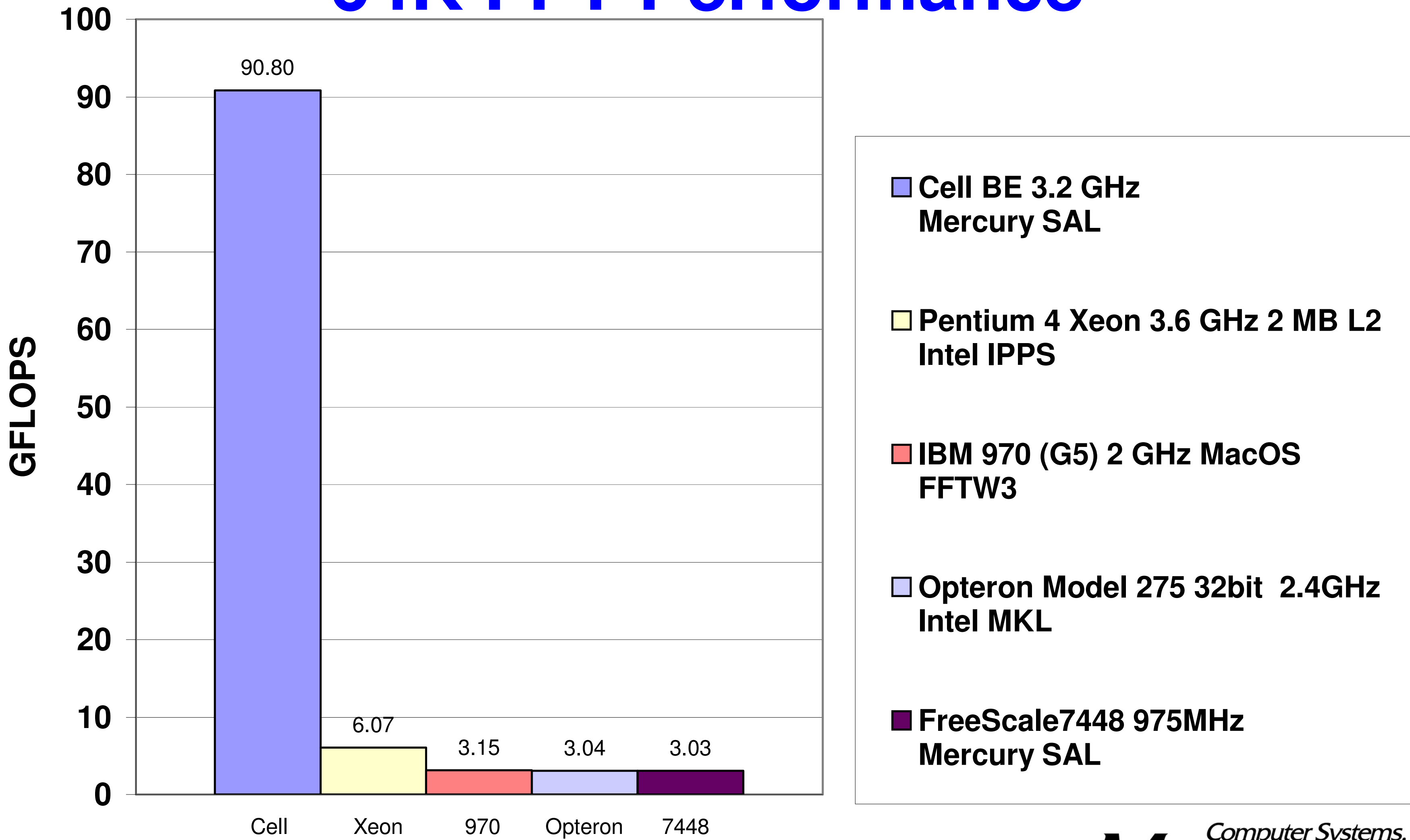
How Much Faster *Is* Cell?



- **Performance relative to 1GHz Freescale 744x (i.e. Freescale = 1)**
- **In all cases, we are comparing Mercury optimized Cell algorithm implementations with the best available (Mercury or 3rd party) implementations on other processors**
- **Did not compare with dual core x86 processors**

Mercury Computer Systems, Inc.

64K FFT Performance



64K Single Precision Complex FFT

Mercury Computer Systems, Inc.

Why is Cell so Fast?

- Between 15 and 30 times faster than comparable GPPs for this algorithm
- Huge inter-SPE bandwidth
  - 205 GB/s sustained throughput
- Fast main memory
  - 25.6 GB/s XDR bandwidth
- Predictable DMA latency and throughput
  - DMA traffic has negligible impact on SPE local store bandwidth
  - Easy to overlap data movement with computation
- High performance, low power SPE cores

Mercury Computer Systems, Inc.

Outline

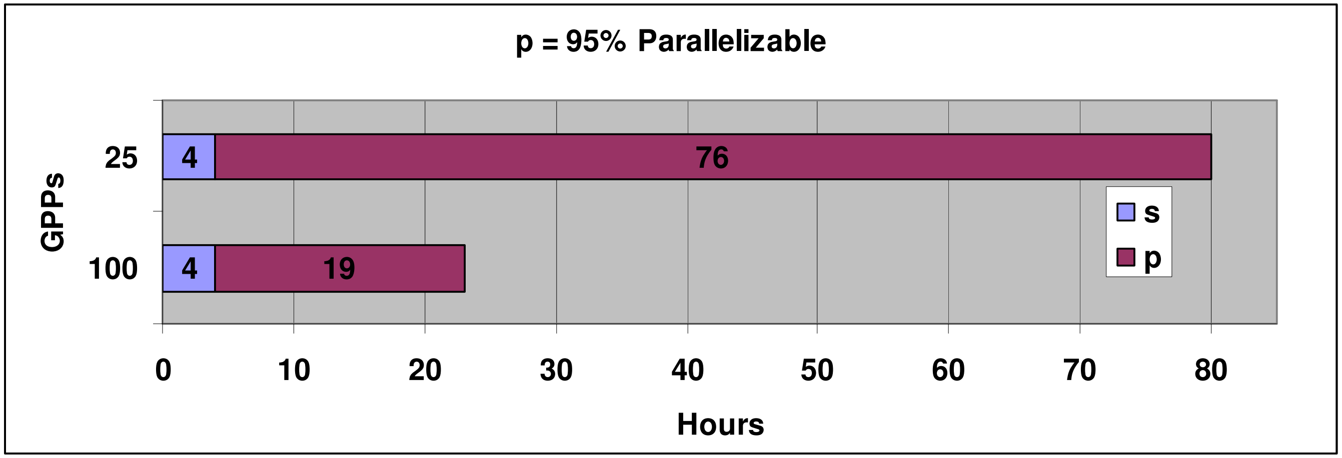
- Introduction
- Distributed Processing Solutions
- Cell Be platform
  - FFT on the Cell
- Calculating Speedup
- Future applications

Amdahl's Law Predicts SCF Speedup

$$Speedup(S) = \frac{N}{s * N + (1 - s)}$$

n = Number of initial processors
N = Number of processors scaled to
s = serialization, or % of job that can't be parallelized
p = parallelization = 1-s

$$Speedup(S) = \frac{1}{1 + p * \left(\frac{n}{N} - 1 \right)}$$

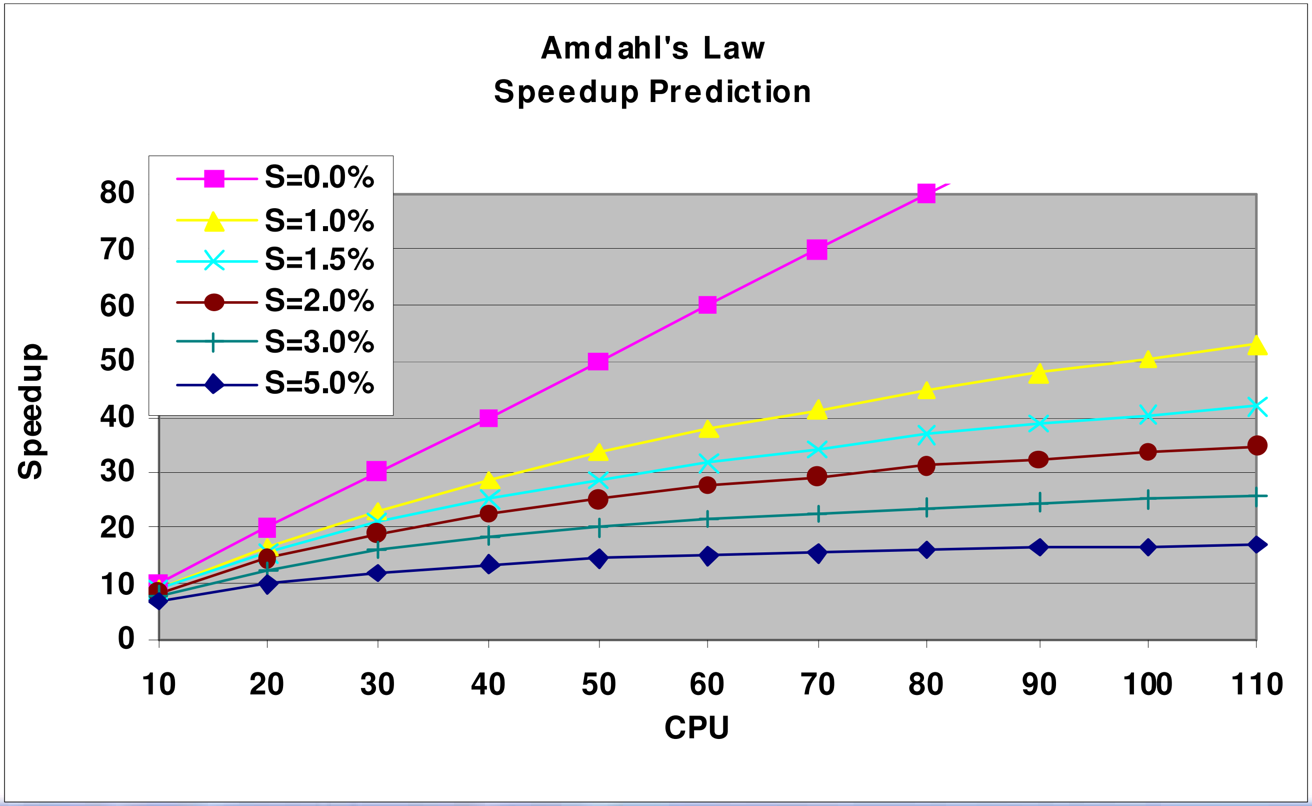


	Baseline	Parallelized
Percent Parallel	95%	95%
Baseline TAT	80	80
Total GPPs (N)	25	100
Sequential Time (s)	4	4
Parallel Time (p)	76	19
Total Sped-Up TAT	80	23
Speedup (S)	3.48	

$$SU = \frac{1}{1 + 0.95 * \left(\frac{25}{100} - 1 \right)} = 3.48$$

p (or s) Determines Scalability

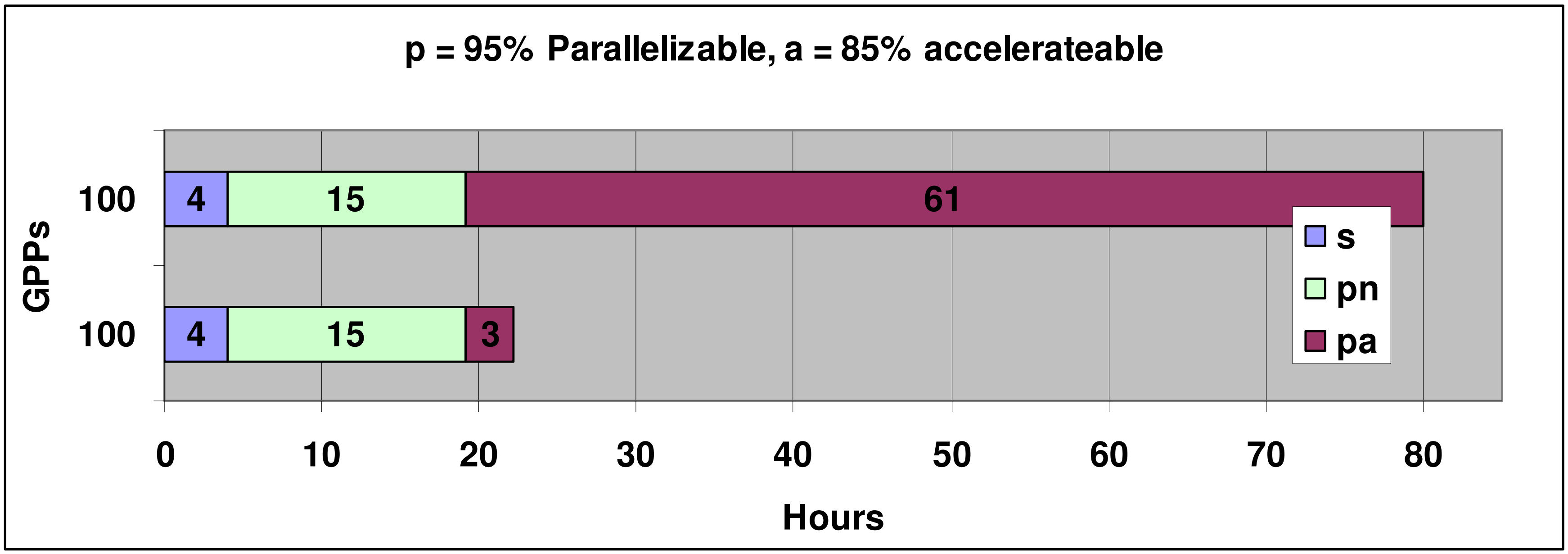
- **The number of processors you can scale, or increase to, and still get useful speedup**



Hybrid Amdahl's Law Predicts HCF Speedup

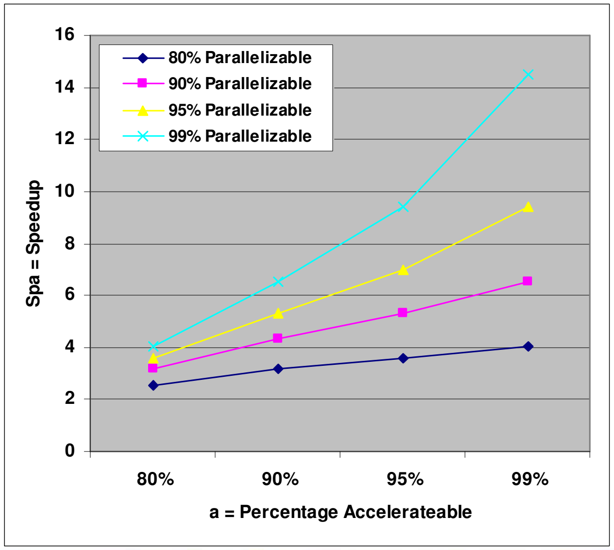
$$Spa = \left\langle \frac{1}{\left\{1 + p\left[\frac{n}{N}\left(1 + a\left(\frac{1}{uA} - 1\right)\right) - 1\right]\right\}} \right\rangle$$

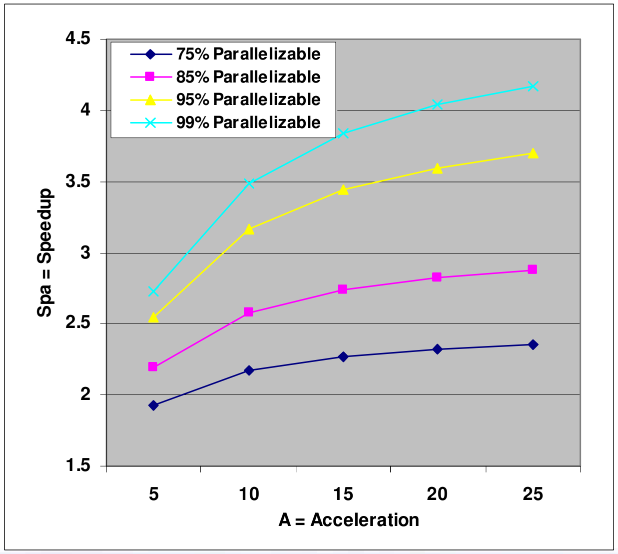
n = Number of initial processors
N = Number of processors scaled to
s = serialization, or % of job that can't be parallelized
p = % parallelization = 1-s
a = % of parallelized job that can be accelerated
A = Acceleration of accelerated tasks provided by CPA
u =% Utilization of CPAs



Scalability of HCF

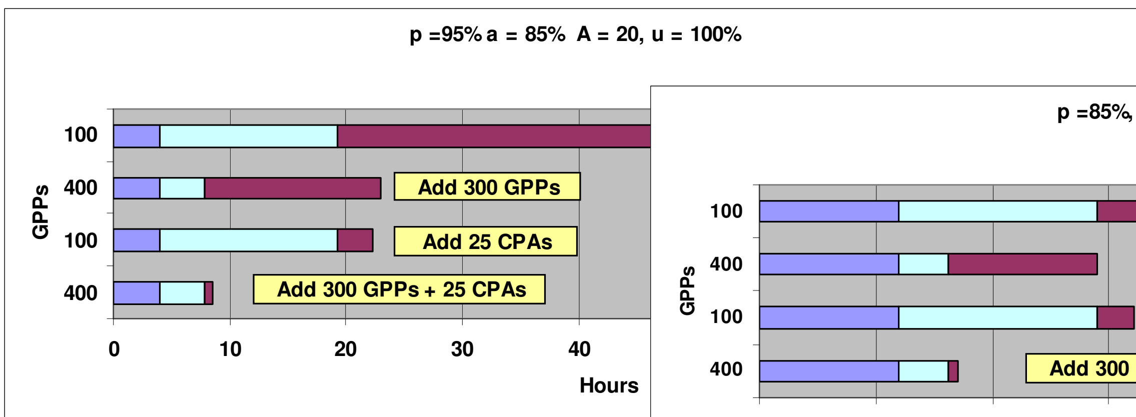
- The higher the parallelization, acceleration percentage, and acceleration, the greater the overall speedup, Spa.
- Spa increases geometrically with both the p factor and the a factor.
- Spa rolls off as A approaches 30, though the roll-off decreases with increasing parallelization. The extent of this roll-off depends on the percentage of the parallelized tasks that can be accelerated, a.

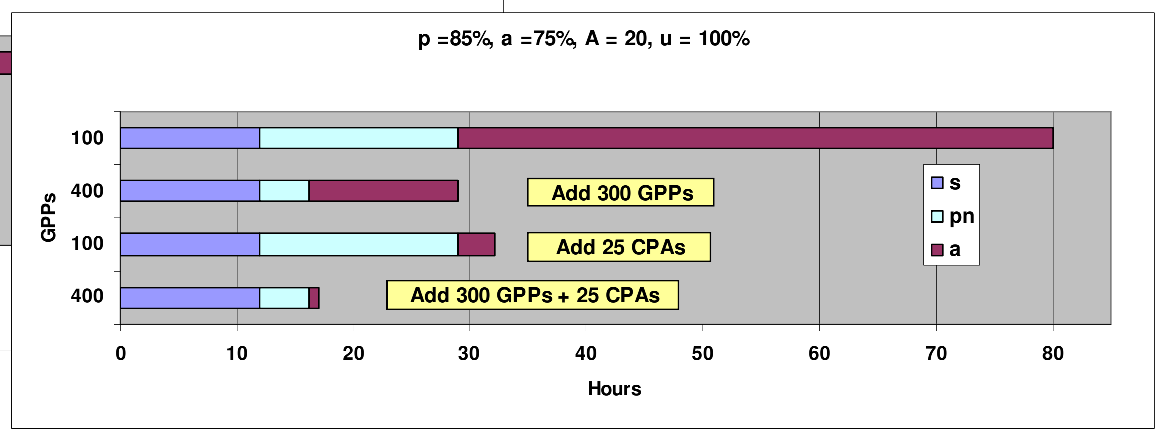




Adding GPPs vs CPAs

	Baseline	Add GPPs	Add CPAs	Add Both
Baseline TAT	80	80	80	80
Basline Servers (n)	100	100	100	100
Additional Servers	0	300	0	300
Parallelizable (p)	95%	95%	95%	95%
CPA Utilization (u)	100%	100%	100%	100%
Acceleratable (a)	80%	80%	80%	80%
CPA Acceleration (A)	1.0	1.0	20.0	20.0
Total Servers (N)	100	400	100	400
Sequential Time	4	4	4	4
Parallel Non-Accelerable Time	15	4	15	4
Parallel Accelerable Time	61	15	3	1
Total Run Time	80	23	22	9
Total Speedup	1.0	3.5	3.6	9.3

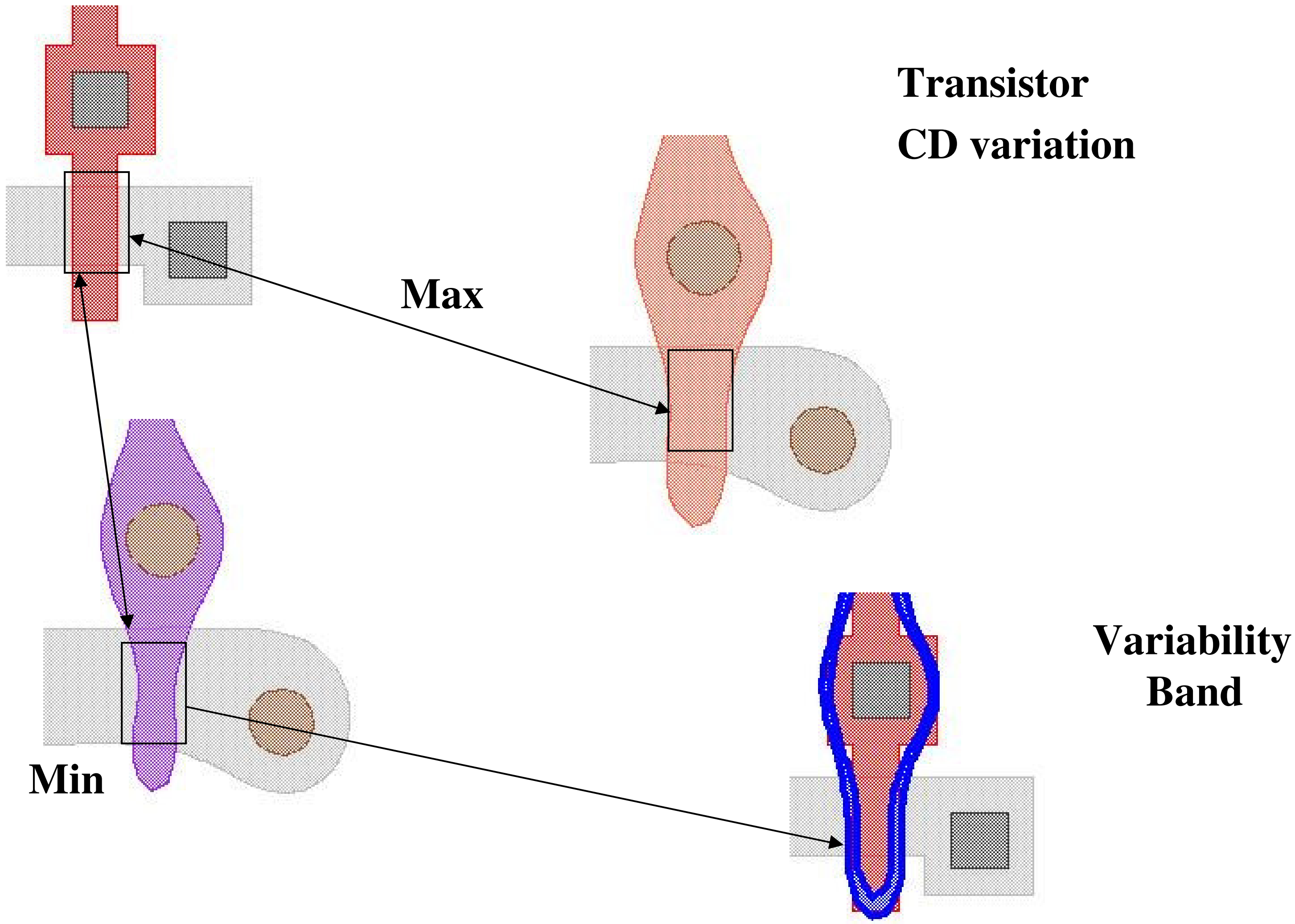




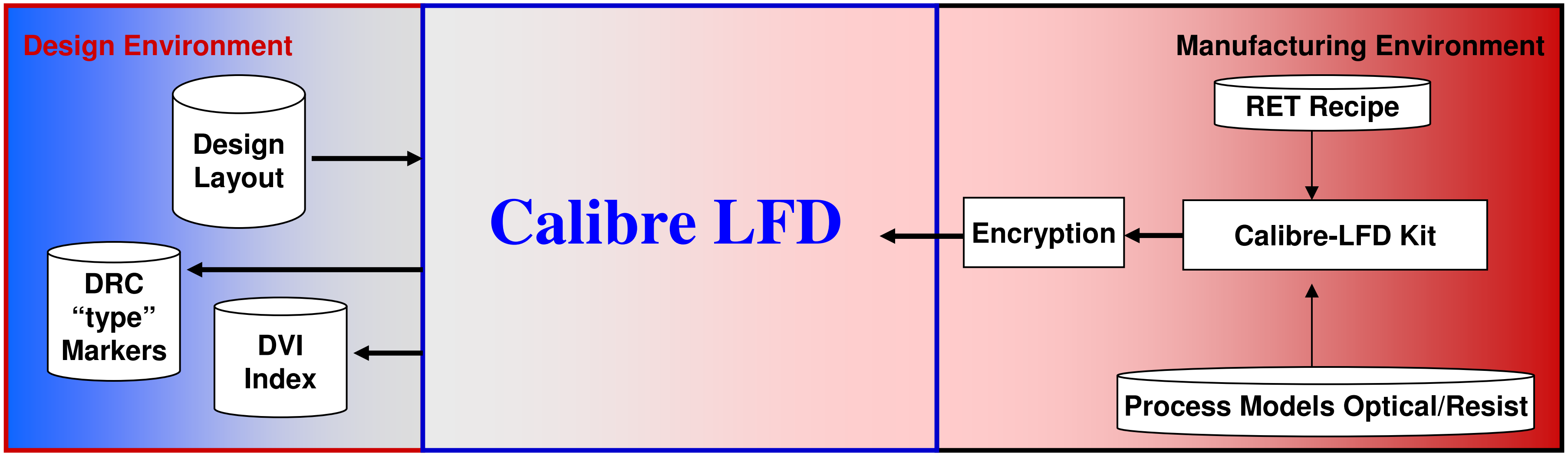
Outline

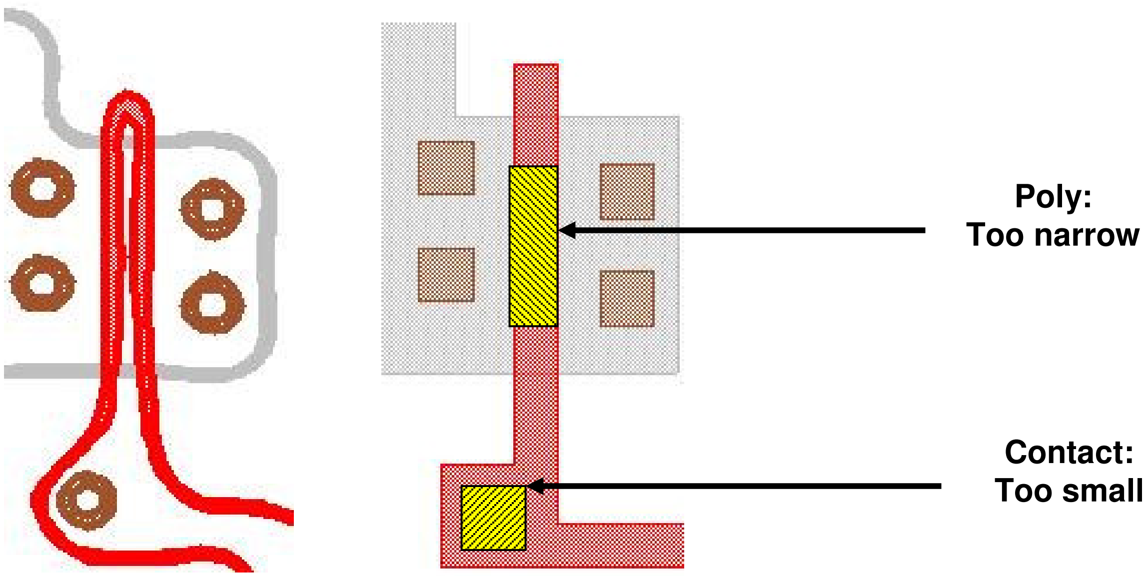
- Introduction
- Distributed Processing Solutions
- Cell Be platform
  - FFT on the Cell
- Calculating Speedup
- Future applications

Power/Leakage variability



Calibre LFD Technology





Layer	DVI
POLY	0.25
CONTACT	0.40

DVI™:
Design Variability Index
Helps designers identify sensitive topologies

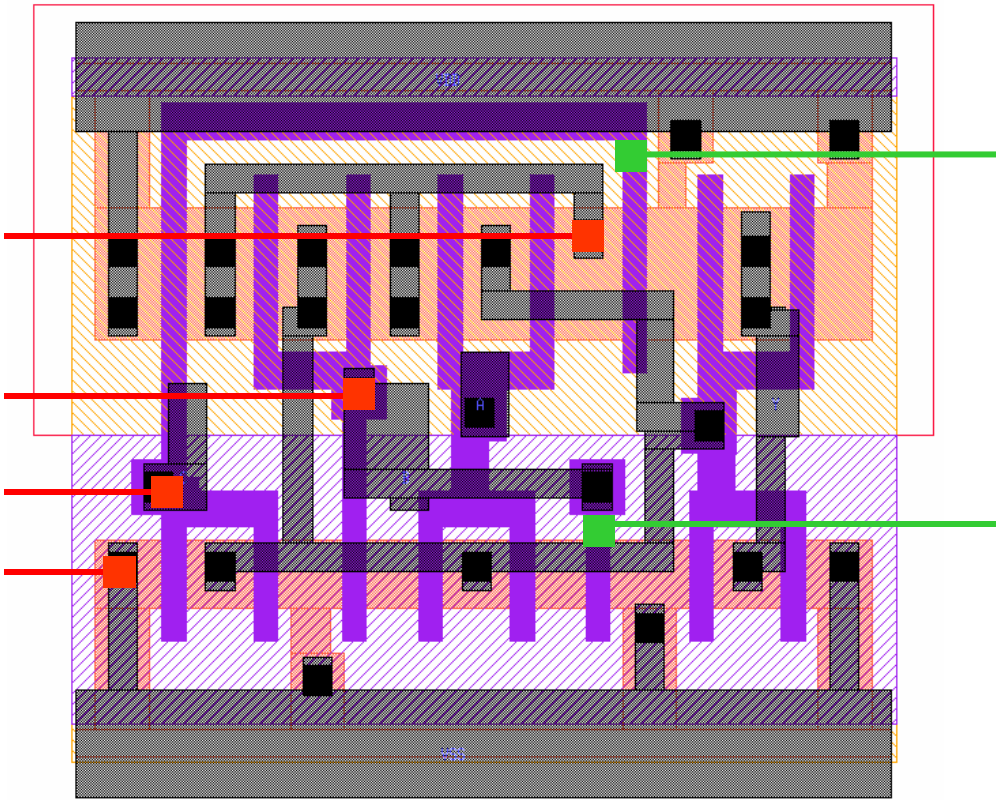
LFD Guided Optimization Example

Layer	DVI
POLY	0.008
CONTACT	0.294
METAL1	0.004

Critical Errors: 6

4-Non resolving contacts

2-Poly silicon pinching

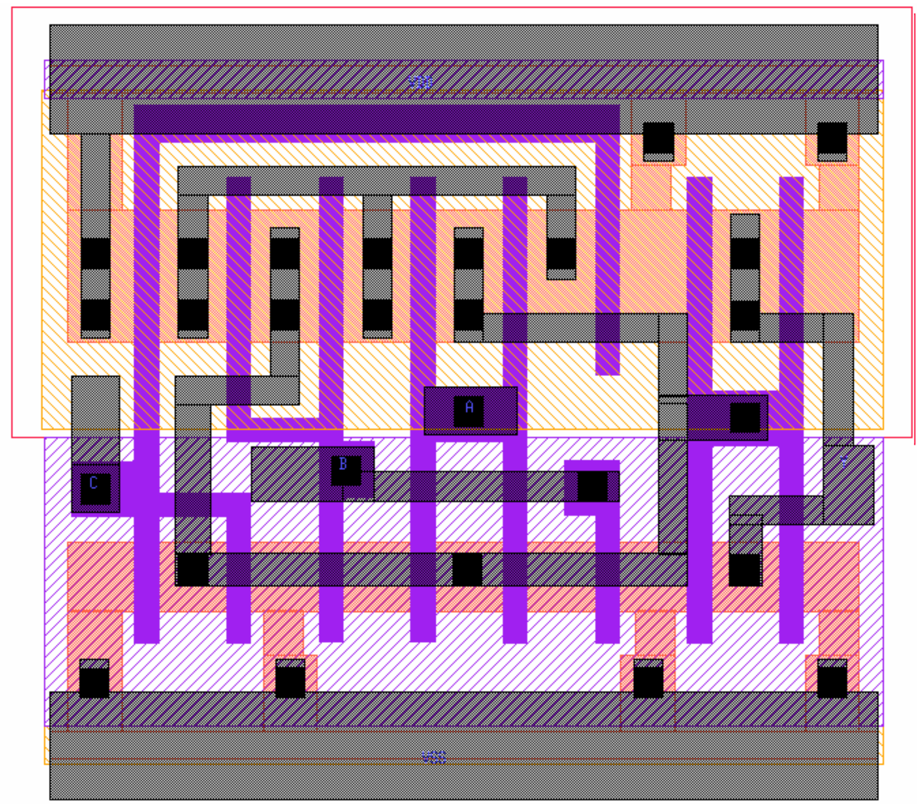


LFD Modifications

Layer	DVI
POLY	0.008
CONTACT	0.294
METAL1	0.004

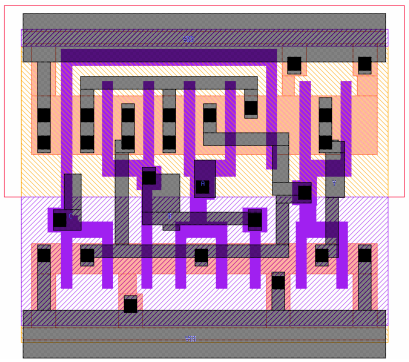
Critical Errors: None

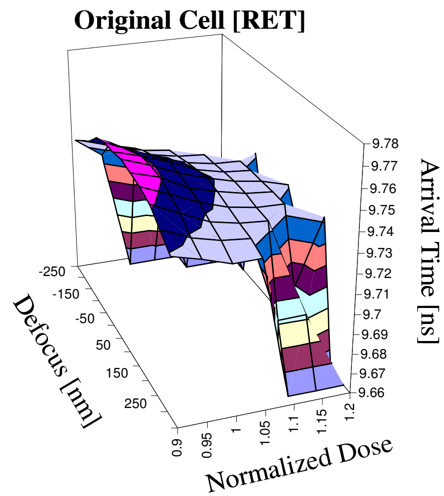
Layer	DVI
POLY	0.007
CONTACT	0.000
METAL1	0.005

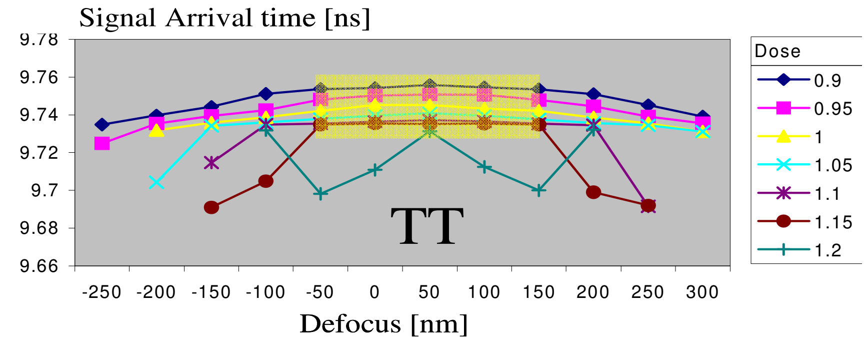


- Place contacts on grid
- Modify metal and poly

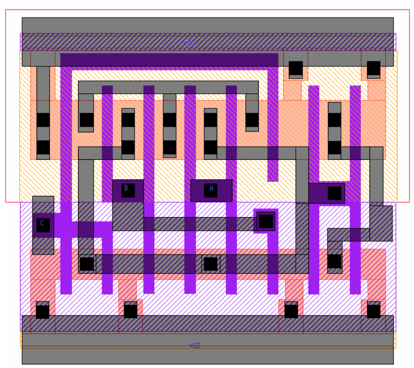
Parametric side-benefit: Timing

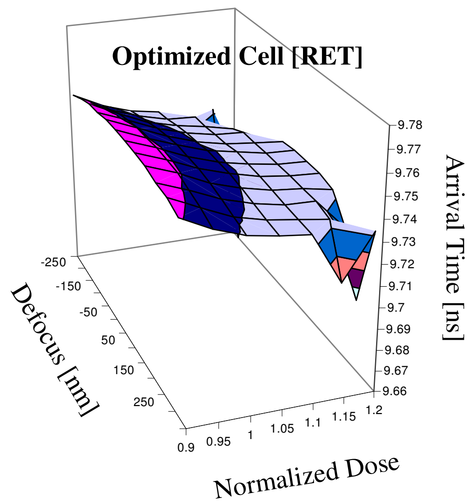


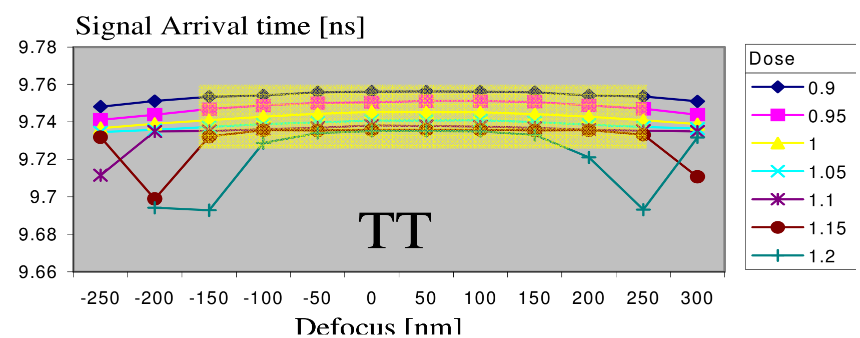




The optimized cell performs within spec even under larger process variability conditions.







Conclusions

- Computational Lithography is growing
  - In more ways than one!
- Flexibility for the job hand becomes critical
  - H/W Platform, Network, OS, software
- Mentor Calibre OPC is working with Mercury
  - Flexible solution for optimal performance
- More to come….

Thank you for your attention.

Mentor Graphics®

www.mentor.com