Energy-Efficient Computing for Embedded Systems

Shanq-Jang Ruan (阮聖彰), Ph. D.
Associate Professor
Dept. Electronic Engineering
National Taiwan University of Science and Technology

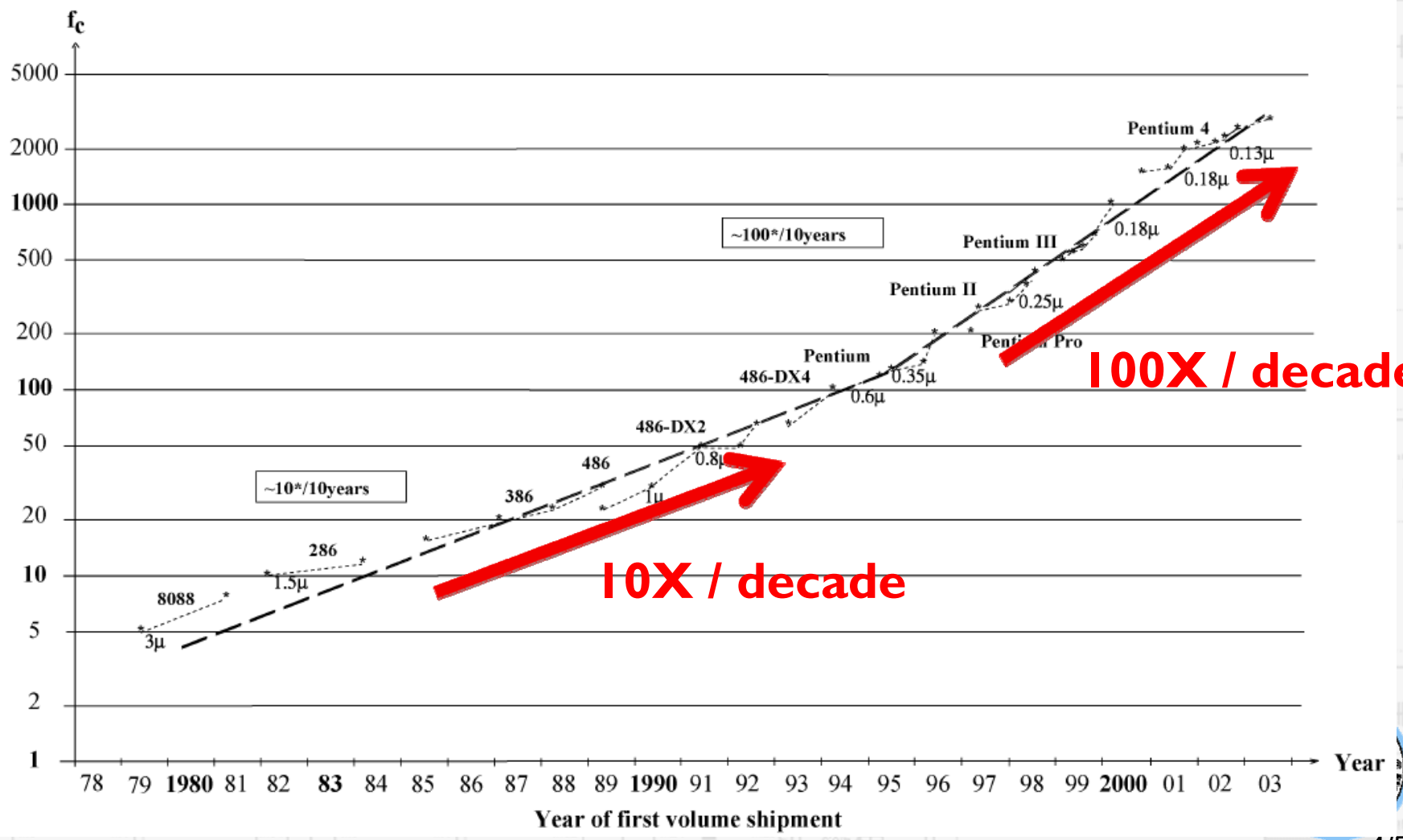
Outline

- Introduction
  - Moore's Law
  - The Power Wall – CPU as an Example
  - Power vs. Performance
  - Embedded Systems
- Power Consumption for Portable Devices
- Low Power Technologies
- Our Recent Research Results
  - A Novel Crosstalk Quantitative Approach for Simultaneously Reducing Power, Noise, and Delay Based on Invert-Base Bus Encoding Schemes
  - Synthesis of Parameter Extractor for Low Power CAM Designs
  - An Efficient Thresholding Algorithm for Document Images Based on Intelligent Block Detection

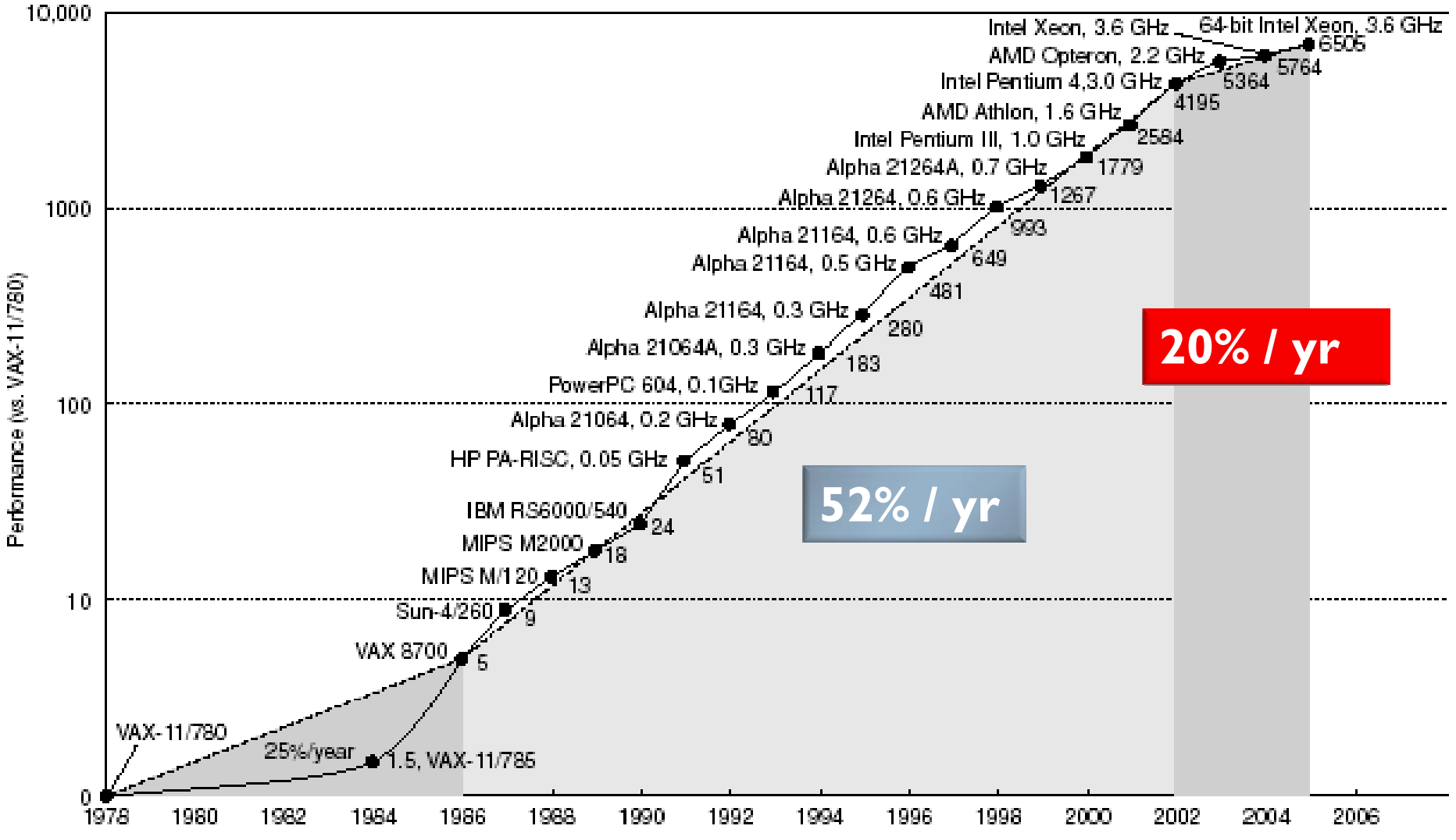
Everything Starts Here

- Moore's Law
  - The concept was introduced in 1965
    - In a paper of Gordon Moore, co-founder of Intel Corp.
  - Double # of trans. every 18~24 months
  - Technology observation and forecast become the goal !!
- However
  - Did not predict performance and power
    - But performance and power grows while transistor scales
    - Will it continue?

Clock Frequency Growth of CPU



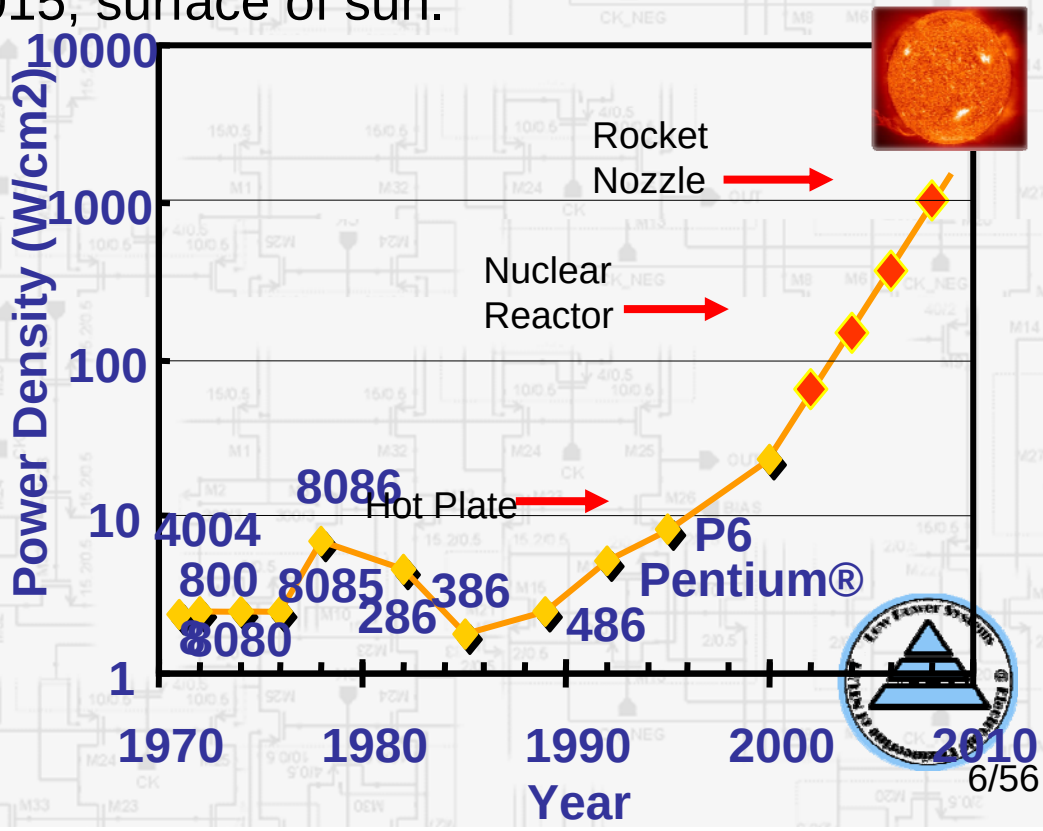
Performance Growth Slows Down



Growth in processor performance since the 1978

Why Worry about Power?

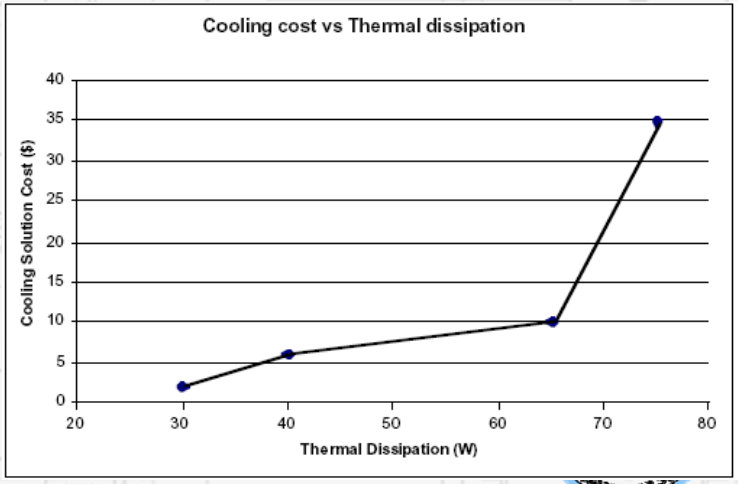
- Intel VP Patrick Gelsinger (ISSCC 2001)
  - If scaling continues at present pace, by 2005, high speed processors would have power density of nuclear reactor, by 2010, a rocket nozzle, and by 2015, surface of sun.
  - "Business as usual will not work in the future."
- Intel stock dropped 8%on the next day.
- But attention to power is increasing.



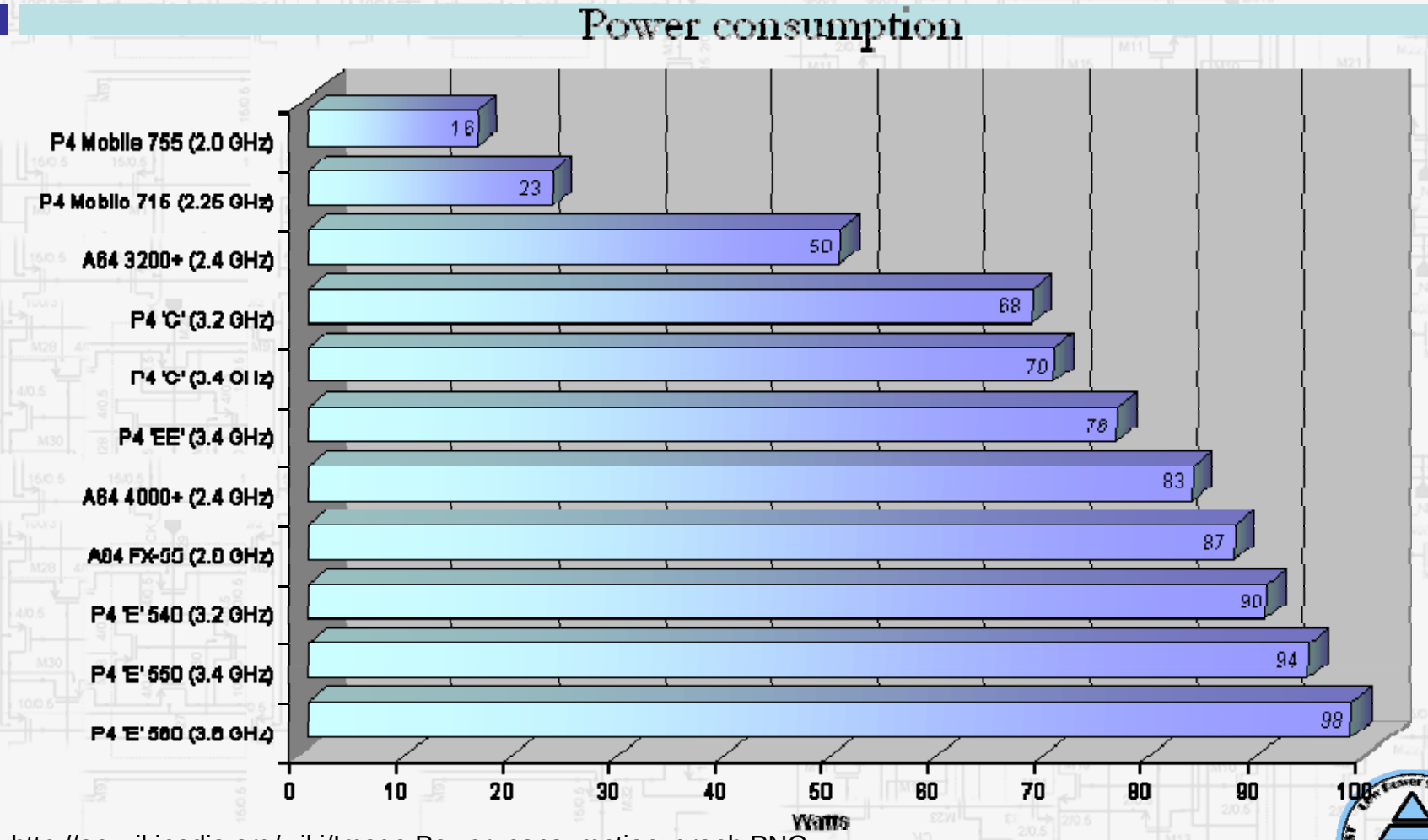
Slide From: Prof. S.J. Ruan, NTUST

Packaging Costs for Heat Dissipation

- Thermal considerations in packing
- The supported temperature for commercial devices during operation equals 0 to 70 ^{o}C.
- Military parts are more demanding and require a temperature range varying from -55 to 125 ^{o}C.
- Increasing the heat removal capacity of a package also tens to raise the package cost.
- The least expensive plastic packaging can dissipate up to 1W.
- Chips dissipating over 20W require special heat sink attachments.

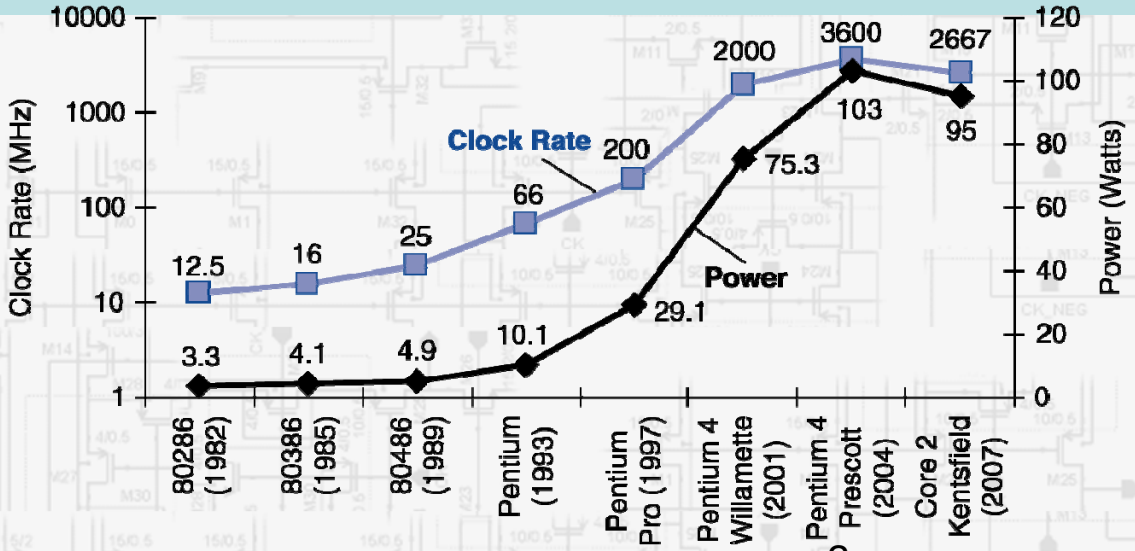


Power Consumption for Intel Processors



http://en.wikipedia.org/wiki/Image:Power_consumption_graph.PNG

The Power Wall



$$\text{Power} = \text{Capacitive load} \times \text{Voltage}^2 \times \text{Frequency}$$

- In 2004, Intel canceled its high-performance uni-processor project (Pentium *4* 4G) because the power problem.
- The future increases in performance will come from more processors per chip rather than higher clock rates and improved CPI.

What Are Multi-Processors?

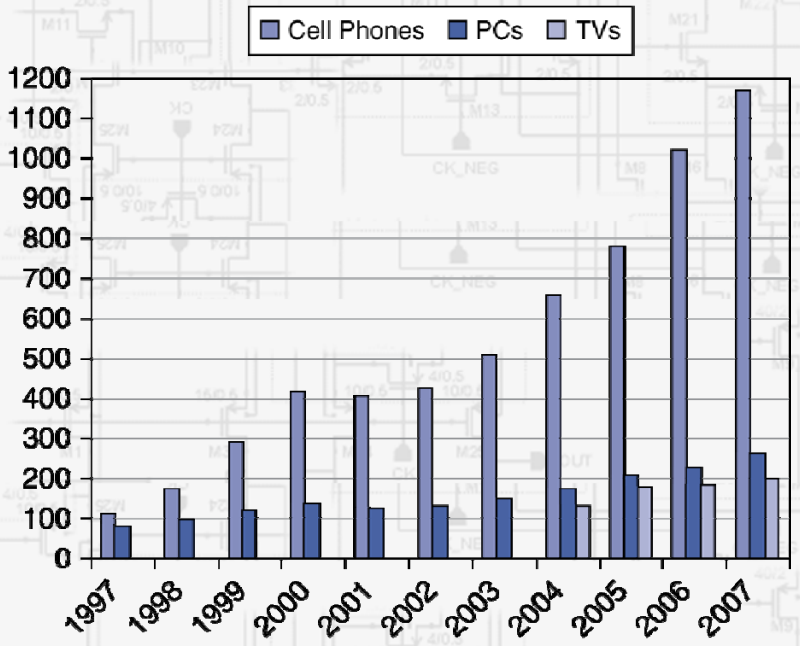
A Single Super Hero

A Well Organized Team

- Present a way to exploit higher parallelism of applications

Embedded Systems

- Embedded computers are the fastest growing portion of the computer market.
- Embedded systems is to meet the performance requirement at a minimum price, rather than achieving higher performance at a higher price.



- Embedded applications: the need to minimize memory and the need to minimize power.
  - Lower power means longer battery life time
  - The need to use less expensive packaging and the absence of a fan for cooling.

Power Consumption for Portable Devices

- What is “Portable Devices”?
- Where does the Power Go in your Notebook
- The Power Dissipation for Portable Devices
- Power Needs vs. Battery Capacities

Multimedia Portable Devices

Example Evolution of Mobile Communications	1G	2G	3G	4G
Increasing Algorithmic Content	Analog Filters	Digital baseband, Voice codec	Digital baseband Voice codec Digital radi Vide	eo codec Camera/Imaging Multi-antenna Multimedia platform GPS Sensors
Higher Performance	Little or	partial IC + Firmware	5Mhz, full ASIC implementation	100Mhz, full ASIC implementation
Integration & Verification Complexity	Analog board simulation	Digital+FW validation	Digital+FW+RF+ Video+SW validation	Digital+FW+ Multichannel RF + GPS+Sensor+ Video +SW validation

Power increasing with each new generation

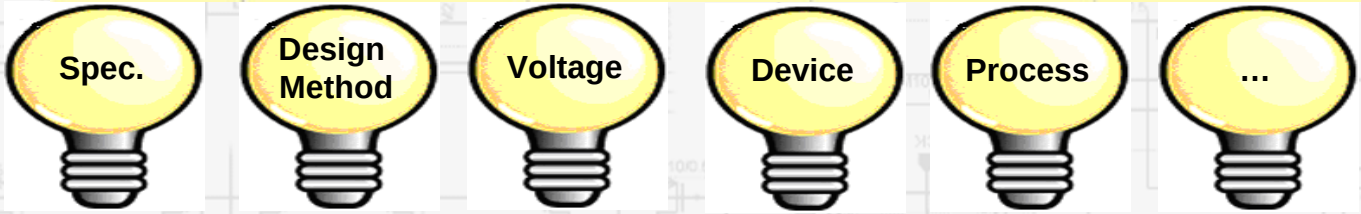
Power Consumption for Portable Devices

- What is “Portable Devices”
  - General definition
    - Mobile devices
  - Specific definition
    - The main power source of mobile devices comes from battery.

It’s all about battery life

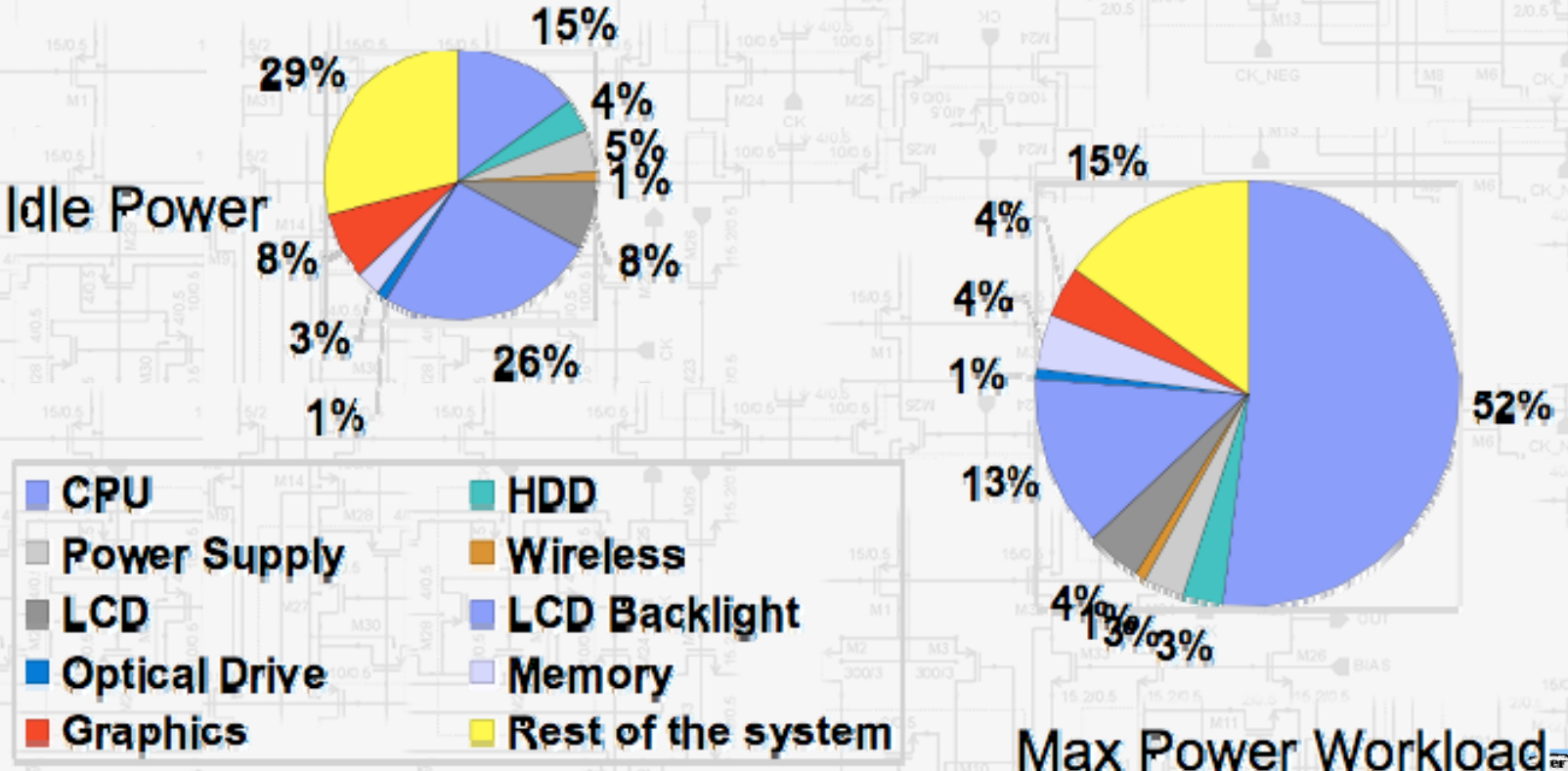
Power Consumption of Mobile Phone

Mobile Phone Benchmark	 Nokia N95 8G	 Nokia N96	 HTC Touch HD	 Apple iPhone 3G 16G
⊟ 使用系統				
型號	Nokia N95-8GB	Nokia N96	HTC Touch HD	Apple iPhone 3G 16GB
上市時間	2007-11-03	2008-09-12	2008-11-28	2008-12-14
市價	$16400	$21800	$23500	$27900
網路頻段	GSM 四頻 / WCDMA	GSM 四頻 / WCDMA	GSM 四頻 / WCDMA	GSM 四頻 / WCDMA
尺寸	99 x 53 x 21 mm	103 x 55 x 18 mm	115 x 62.8 x 12 mm	115.5 x 62.1 x 12.3 mm
主螢幕	240 x 320 pixel 1670 萬色 2.8 吋 TFT	240 x 320 pixel 1670 萬色 2.8 吋 TFT	480 x 800 pixel 3.8 吋 TFT	480 x 320 pixel 3.5 吋 TFT
次螢幕	無	無	無	無
重量	128 g	125 g	146.4 g	133 g
通話時間	300 分鐘	220 分鐘	390 分鐘	600 分鐘
待機時間	280 小時	220 小時	450 小時	300 小時
標準電池	1200 mAh	950 mAh	1350 mAh	未知



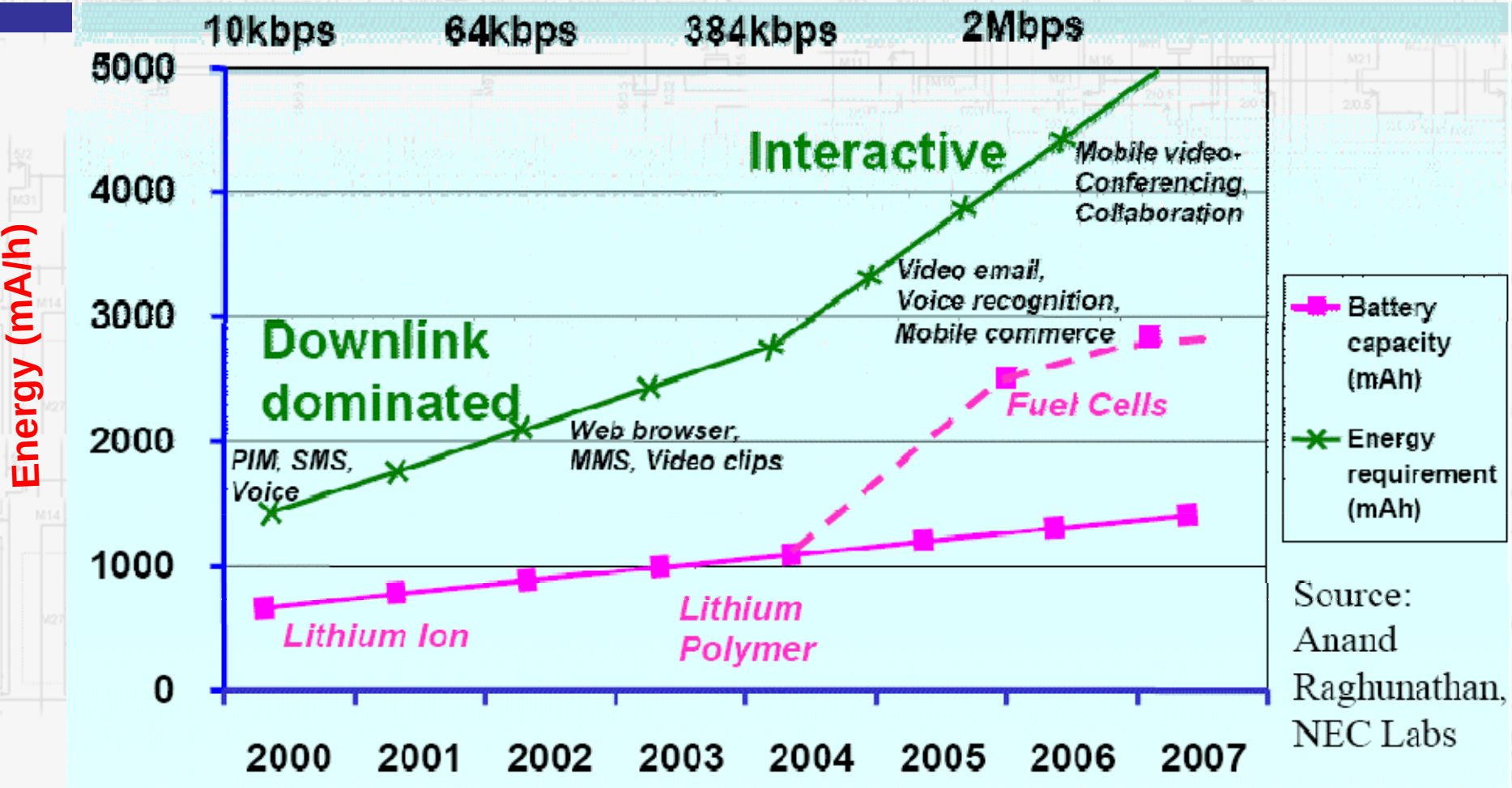
Where Does the Power Go in Your NB?

Current Generation Laptop Power Pie (IBM Thinkpad R40)



Data courtesy Mahesri et al., U of Illinois, 2004

Power Needs vs. Battery Capacities



Source: Anand Raghunathan, NEC Labs

Low Power Technologies

- Power and Energy
- Power Management Spectrum
- Hardware Low Power mechanisms
- Software Low Power techniques
- Specialized Processing Elements

Low Power Technologies

- Power and Energy
- Power Management Spectrum
- Hardware Low Power mechanisms
- Software Low Power techniques
- Specialized Processing Elements

Power and Energy

Power and Energy Formula

- Power is drawn from a voltage source attached to the V_{DD} pin(s) of a chip.

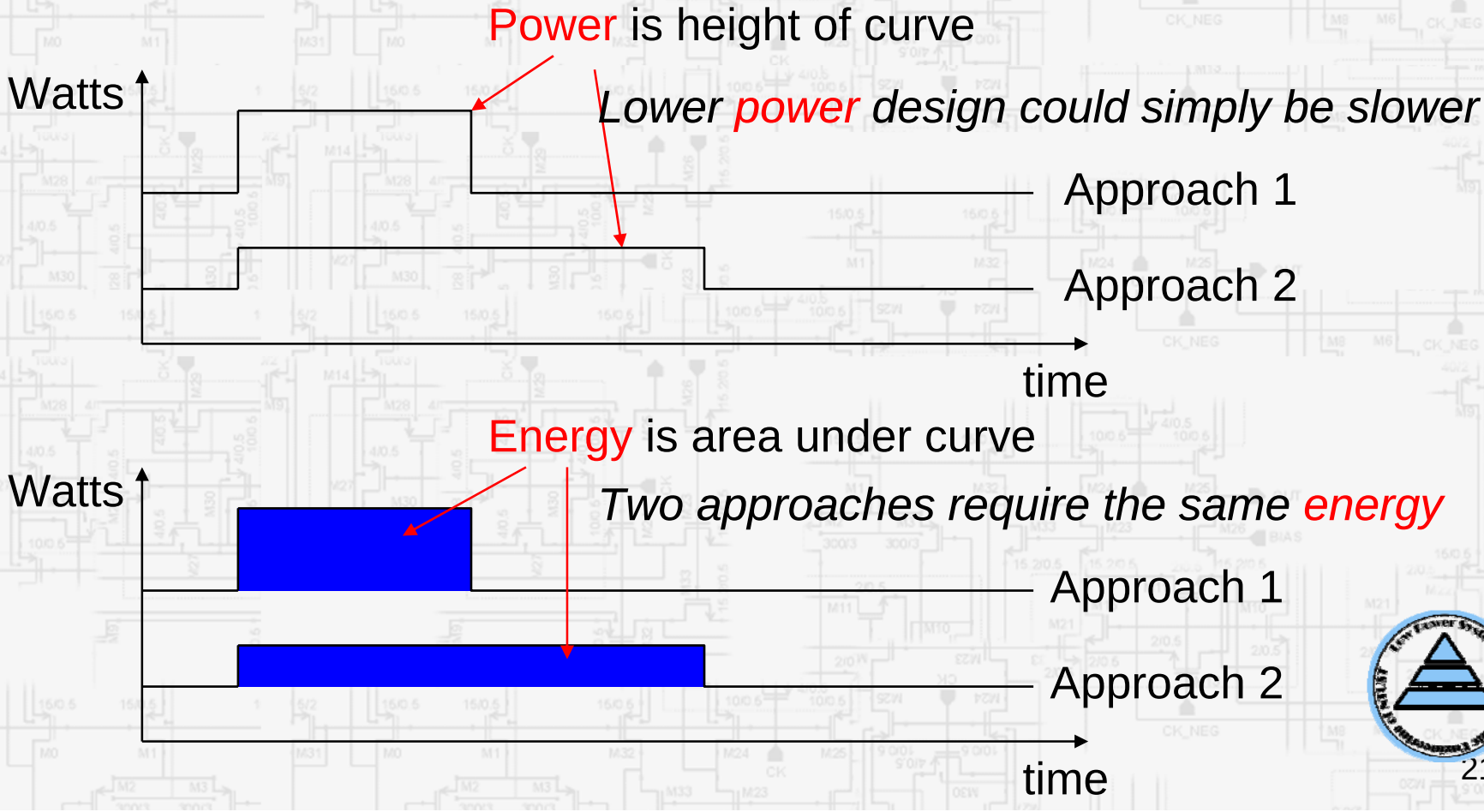
- Instantaneous Power: $$P(t) = i_{DD}(t)V_{DD}$$

- Energy: $$E = \int_0^T P(t)dt = \int_0^T i_{DD}(t)V_{DD}dt$$

- Average Power: $$P_{\text{avg}} = \frac{E}{T} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}dt$$

Power and Energy (cont'd)

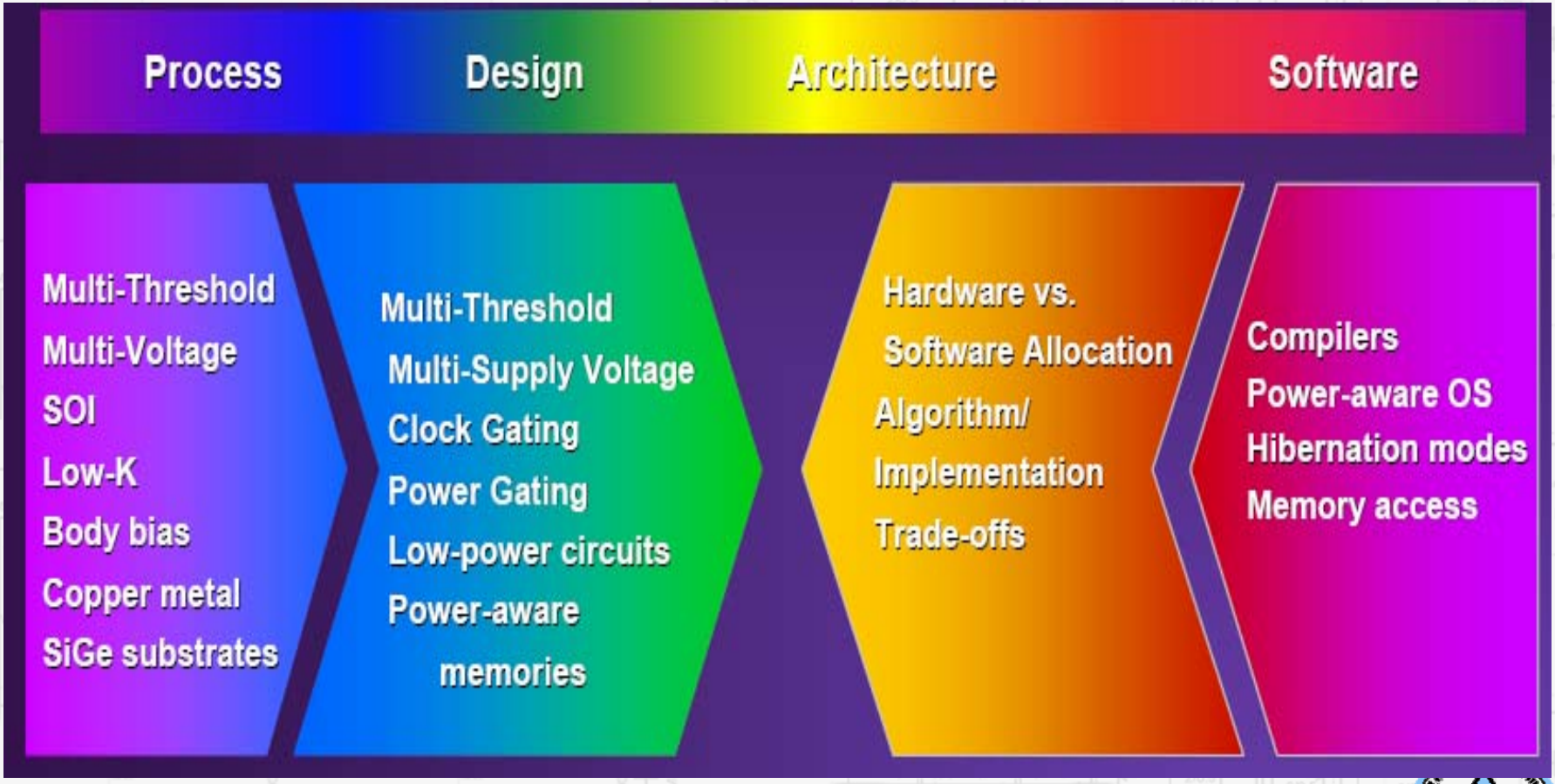
- The relation is between power and energy



Low Power Technologies

- Power and Energy
- Power Management Spectrum
- Hardware Low Power mechanisms
- Software Low Power techniques
- Specialized Processing Elements

Power Management Spectrum



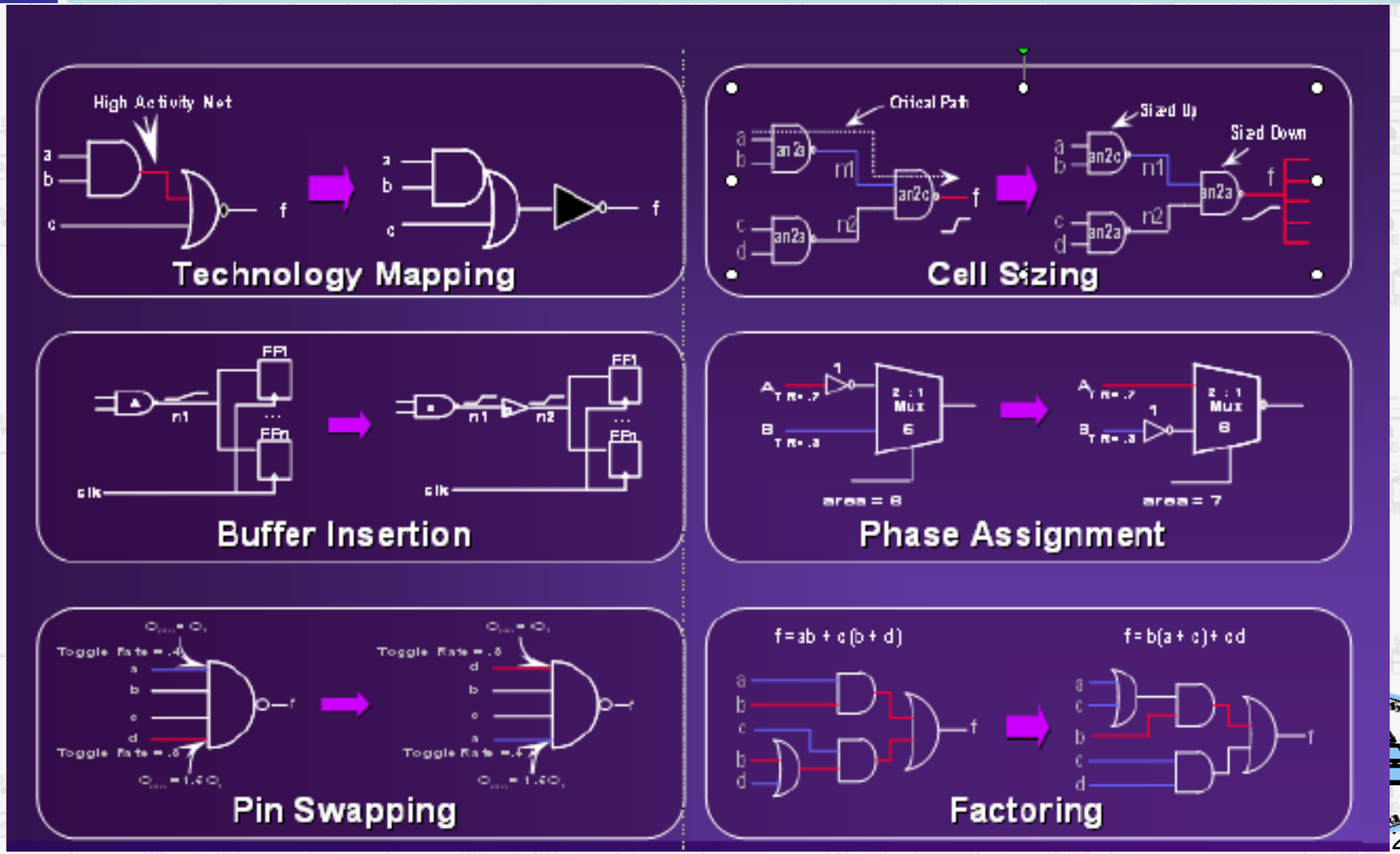
Low Power Technologies

- Power and Energy
- Power Management Spectrum
- Hardware Low Power mechanisms
- Software Low Power techniques
- Specialized Processing Elements

Hardware Low Power Technologies

- Gate Level Optimizations
- Clock Gating
- Reduce Voltage
- Reduce Frequency
- Reduce Components
- Specialized Processing Elements
- Multi-cores and Parallel Technologies

Example: Gate-Level Power Optimizations



Low Power Technologies

- Power and Energy
- Power Management Spectrum
- Hardware Low Power mechanisms
- Software Low Power techniques
- Specialized Processing Elements

Software Low Power Technologies

- Low Power Compiler Optimizations
- DVFS and DPM
- Software Power Management Framework
- Dynamic Voltage and Frequency Scaling

Software Low Power Technologies

- Low Power Compiler Optimizations
- DVFS and DPM
- Software Power Management Framework
- Dynamic Power Management
- Dynamic Voltage and Frequency Scaling

Low Power Compiler Optimizations

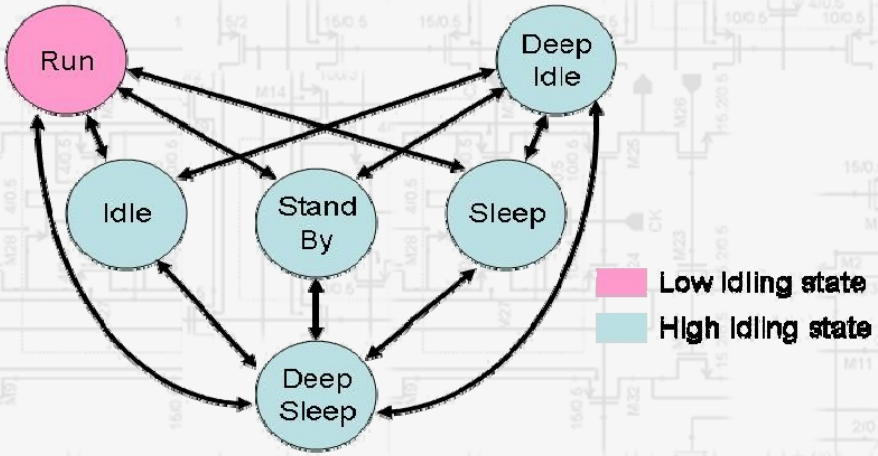
- Redundant instruction elimination
- Induction variable elimination
- Local/global common sub-expression elimination
- Loop unrolling, in-lining, software pipelining
- Loop transformation
- Many others…
- Basically, reducing the instruction count reduces the required energy.

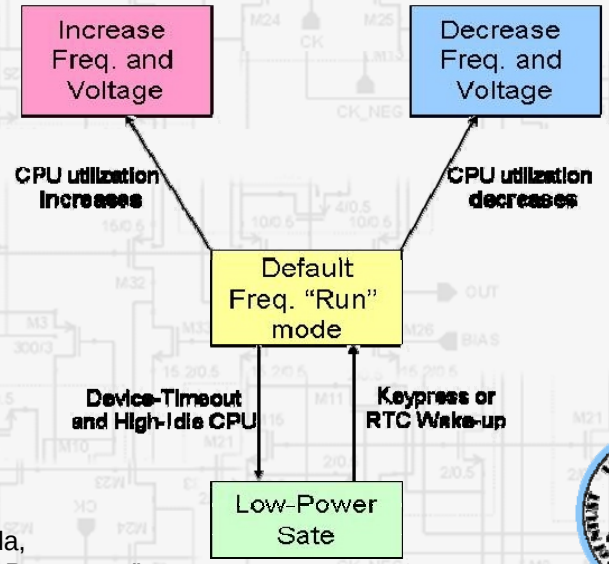
Software Low Power Technologies

- Low Power Compiler Optimizations
- DVFS and DPM
- Software Power Management Framework
- Dynamic Power Management
- Dynamic Voltage and Frequency Scaling

DVFS and DPM

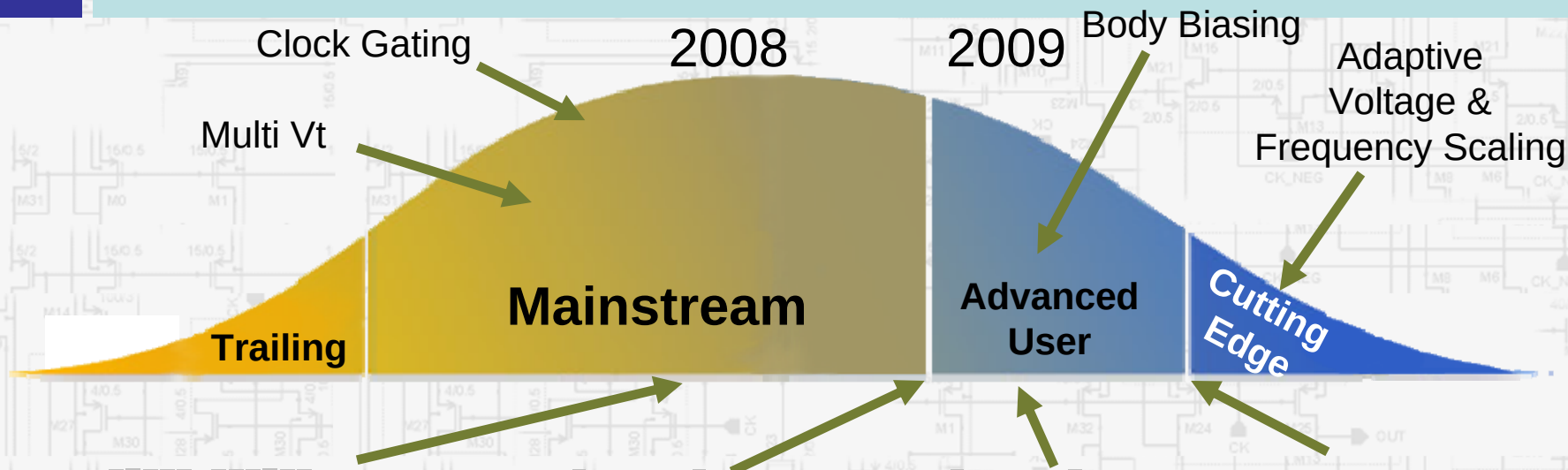
(source: “Wireless Intel SpeedStep Power Manager”, http://www.intel.com/design/pca/applicationsprocessors/whitepapers/300577.htm)

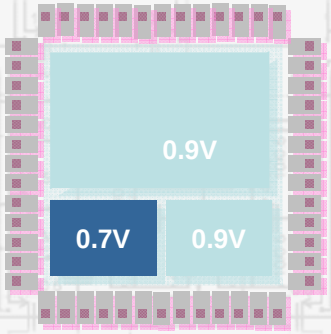




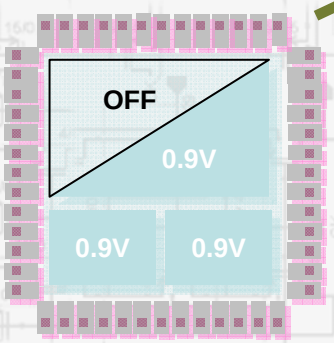
source: Priya N. Vaidyna, Moinul H. Khan, Bryan Morgan, and Premanand Sakarda,
“System Level Adaptive Framework for Power and Performance Scaling On Intel® PXA27X Processor.”

Low Power Technique Adoption

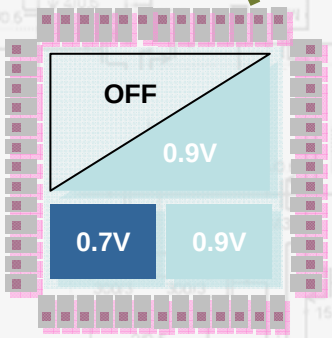




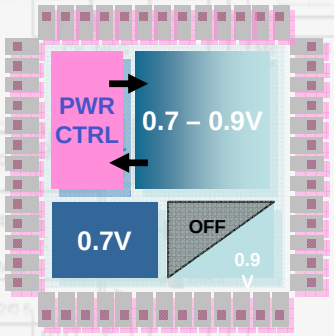
Multi-Voltage (MV)



MTCMOS power gating (shutdown)



MV with power gating



Dynamic Voltage & Frequency Scaling (DVFS)

Software Low Power Technologies

- Low Power Compiler Optimizations
- DVFS and DPM
- Software Power Management Framework
- Dynamic Power Management
- Dynamic Voltage and Frequency Scaling

Dynamic Power Management

- What is DPM?
- Categorization of DPM Techniques
- Options to Reduce Power
- Core Power Management Techniques
- Making the Driver DPM Capable

What is DPM?

- A technology enabling intelligent and dynamic power usage
- Finer-grained control of energy consumption
- Enable devices to enjoy same life with smaller, cheaper power cells
- A DPM system contains a set of rules and procedures to move from one operating point to another as programs execute and events occur.

Power Mode Examples

Lucent's Wavelan PCMCIA Card		Rambus DRAM (by IDF: Mobile Rambus Spec)		Flash RAM: AMD/Fujitsu Am42BDS6408G	
Mode	**Power**	**Mode**	**Power**	**Mode**	**Current**
Transmit	1.82W	Active	300 mW	Simultaneous Operation	25 mA
Receive	1.80W	StandBy	180 mW	Program/Erase	15 mA
StandBy	0.18W	Nap	30 mW	Burst Read	10 mA
		PowerDown	3 mW	StandBy	0.2 µA

Categorization of DPM Techniques

- Timeout policy
  - Shutting down unused devices by a static timeout threshold
- Predictive technique
  - According to the workload history, decides when to shut down unused devices
- Stochastic process
  - According to a random process model, decides when to shut down unused devices

Options to Reduce Power

- Optimize power usage when active, e.g.
  - Scale screen output
  - One chip/core solution vs. more
  - Reduce memory usage
  - Scale processor and component speed/voltage
- Minimize power usage when inactive, e.g.
  - Optimize start-up and shutdown times
  - Reduce unnecessary system maintenance activities
  - Shutdown power to as many components as possible

Based on documents from Montavista

Options to Reduce Power (cont'd)

Component	Power Management Opportunities
CPU Core	Clock frequency and voltage; sleep/halt
DSP Core	Clock frequency and voltage, disable
RAM	Bus clock, DRAM refresh rate, burst modes, enable/disable
Flash	Access modes, write enable/disable, use of shadow RAM
System Bus	Vary voltage and frequency
LCD Display	Change scan frequency, turn off/dim backlight, power-down
Audio Output	Scale volume, disable output stage
Base band/RF	Cycle power, disable RF output stage

Based on documents from Montavista

Software Low Power Technologies

- Low Power Compiler Optimizations
- DVFS and DPM
- Software Power Management Framework
- Dynamic Power Management
- Dynamic Voltage and Frequency Scaling

Dynamic Power and DVFS

$$P_{dyn} = N_{sw} C_L V_{DD}^{\ 2} f$$

- N_{sw}: the switching activity
- C_L: the load capacitance
- V_{DD}: the supply voltage
- f : the operation frequency

$$delay = \frac{1}{f} = k \times \frac{V_{DD}}{(V_{DD} - V_{th})^{\alpha}}$$

- k: the process and gate size
- V_{th}: the threshold voltage
- α : the channel constant (1~2)

The Supporting frequency list of the Creator PXA270 platform

CPU Frequency	Bus Frequency	Mem. Frequency	Min. CPU Voltage
104 MHz	208 MHz	104MHz	0.9V
156 MHz	104 MHz	104MHz	1.0V
208 MHz	208 MHz	104MHz	1.15V
312 MHz	208 MHz	104MHz	1.25V
416 MHz	208 MHz	104MHz	1.35V
520 MHz	208 MHz	104MHz	1.45V

Example Processors supporting DVFS

Intel Pentium !!!(M) Ultra-low power version		AMD Mobile Athlon 4		Transmeta Crusoe TM5500	
Freq.	Volt.	Freq.	Volt.	Freq.	Volt.
500MHz	1.1	500MHz	1.2	300MHz	0.9
600MHz	1.1	600MHz	1.25	400MHz	1.0
650MHz	1.6			533MHz	1.1
700MHz	1.6	700MHz	1.3	667MHz	1.2
750MHz	1.35				
750MHz	1.6				
800MHz	1.6	800MHz	1.35	800MHz	1.3
850MHz	1.6				
900MHz	1.7				
1 GHz	1.7	1 GHz	1.4		

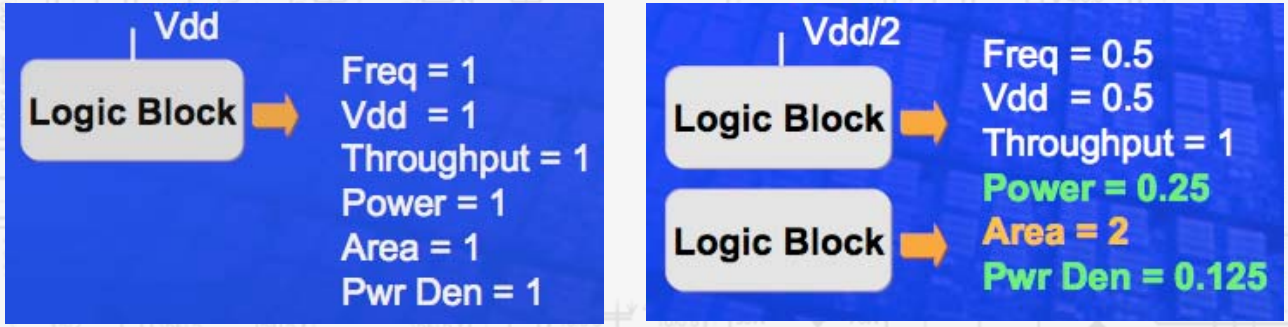
Multi-Core Low Power Technologies

- Multi-cores and parallel execution
  - Reduce frequency of each processor
  - Achieve the original performance by parallel execution over multiple processors
- Review of the dynamic power equation

$$P = 0.5 \times C_L \times V_{dd}^2 \times E(sw)$$

Multi-Core Low Power Technologies (cont'd)

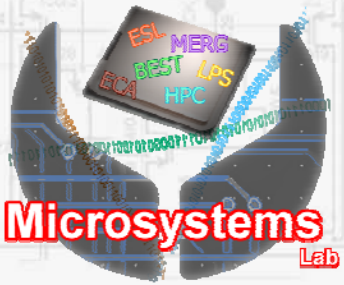
- Example:
  - Single core: 3GHz, 2VDC
    - Power consumption P ~ K*C*F*V^2
  - Dual cores: 1.5 GHz, 1VDC
    - Ideal Power consumption P' ~ K*(2C)*(½ F)*(½ V)^2 = ¼ P
      - 2C: 2 cores



Low Power Through Specialized Processing Elements or Hardware

- Parallel hardware
- DSP processors
- Coprocessors
- Codec
- FPGA
- Reconfigurable processors

Examples



Our Recent Research Results

TAIWAN TECH
National Taiwan University of Science and Technology

A Novel Crosstalk Quantitative Approach for Simultaneously Reducing Power, Noise, and Delay Based on Invert-Base Bus Encoding Schemes

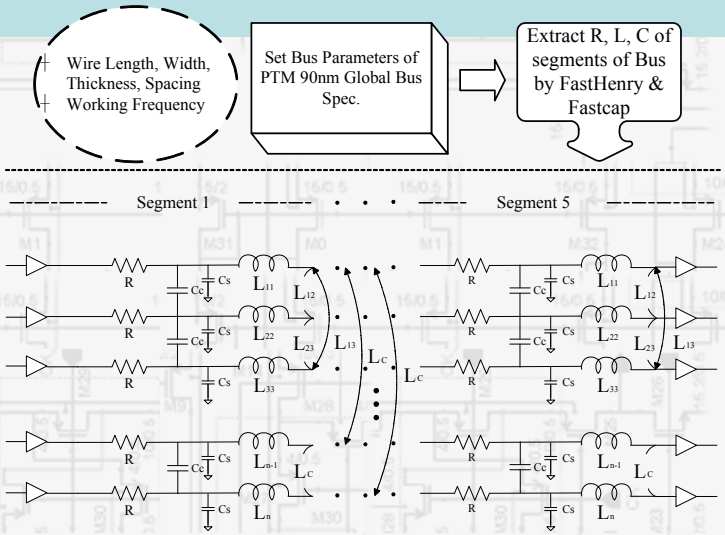
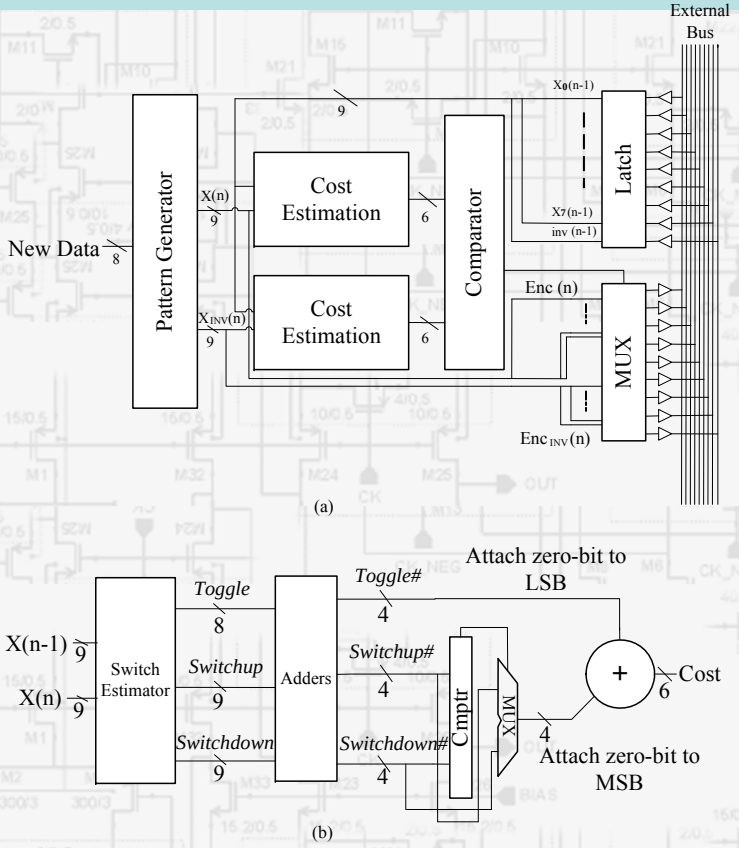


Table 1. Variations of power and noise with different pairs of transition patterns (Vdd = 1.0V)

Switching Pattern Pair	Power Consumption	Max Swing Voltage (V)
↑↓------	11.53	0.99
--↑↑↑↑↑	3.75	1.49
↑↓↑-----	19.31	1.32
---↑↑↑↑	3.42	1.45
↑↓↑↓----	26.8	1.16
----↑↑↑↑	3.21	1.38
↑↓↑↓↑---	34.29	1.19
-----↑↑↑	3.04	1.33
↑↓↑↓↑↓--	41.95	1.20
------↑↑	2.82	1.23



(a) Encoder architecture (b) Cost estimation

$$Cost = Toggle\# \times 2 + MAX(Switchup\#, Switchdown\#).$$

A Novel Crosstalk Quantitative Approach for Simultaneously Reducing Power, Noise, and Delay Based on Invert-Base Bus Encoding Schemes

Table 2. Peak and average power and delay reduction of different bus widths.

Bus Width	Peak P_{org_bus}	Peak P_{enc_bus}	AVG P_{org_bus}	AVG P_{enc_bus}	Overhead P_{codec}	Total Peak Power ($P_{enc\_bus}+P_{codec}$)	Peak PR(%)	AVG PR(%)	$Delay_{org}$ (ps)	$Delay_{enc}$ (ps)	DR (%)
8	95.29	52.08	36.65	28.23	1.1	53.18	44.19	19.97	322	265	19.7
16	141.46	91.9	70.2	57.62	1.85	93.75	33.73	15.29	402	350	12.94
32	272.81	194.8	135.98	116.65	4.32	199.12	27.01	11.04	456	413	9.43
64	520.29	371.56	271.83	232.95	8.6	380.16	26.93	11.13	498	463	7.03

Table 3. Peak and average noise reduction of different bus widths.

Bus Width	Peak Norg	Peak Nenc	Peak NR(%)	AVG Norg	AVG Nenc	AVG NR(%)
8	1.78	1.47	17.42	1.29	1.1	14.73
16	1.93	1.68	12.95	1.42	1.22	14.09
32	2.14	1.89	11.68	1.53	1.32	13.73
64	2.18	1.96	10.09	1.56	1.38	11.54

Table 4. Comparison of power, noise, and delay reduction of an 8-bit bus by using different methods.

Encoding Scheme	Peak Power	Peak Noise	AVG Power	AVG Noise	Codec Power	Peak PR(%)	AVG PR(%)	Peak NR(%)	AVG NR(%)	Delay (ps)	DR (%)
Un-coded data	95.29	1.78	36.65	1.29	-	-	-	-	-	322	-
Bus-Invert method [8]	69.31	1.44	32.2	1.09	0.65	26.58	10.37	19.1	15.5	256	20.5
Tu's method [7]	82.79	1.42	34.1	1.07	0.76	12.32	4.88	20.22	17.05	279	13.35
Our Proposed Scheme	52.08	1.47	28.23	1.1	1.1	44.19	19.97	17.42	14.73	265	17.7

Synthesis of Parameter Extractor for Low Power CAM Designs

Easy, Fast, Identification

Small

Lowering comparisons

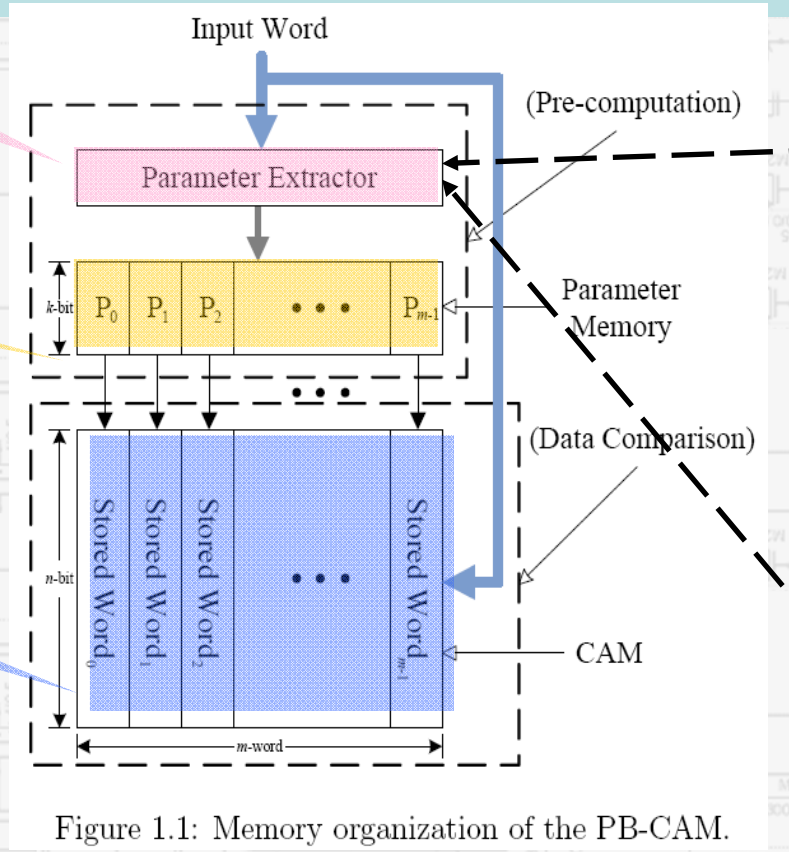


Figure 1.1: Memory organization of the PB-CAM.

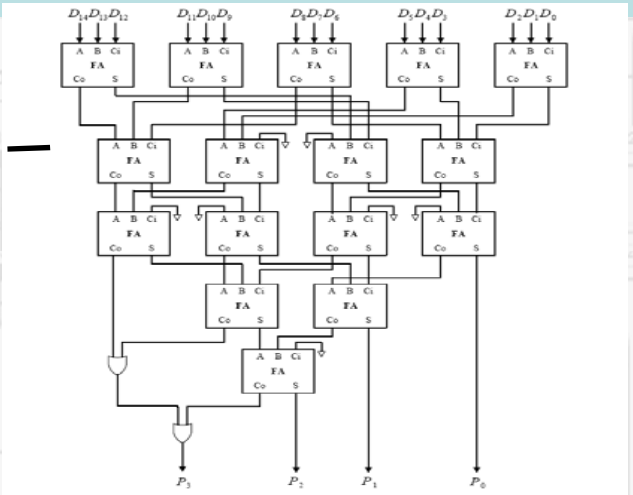


Figure 1.4: 15-bit parameter extractor of the 1's count PB-CAM.

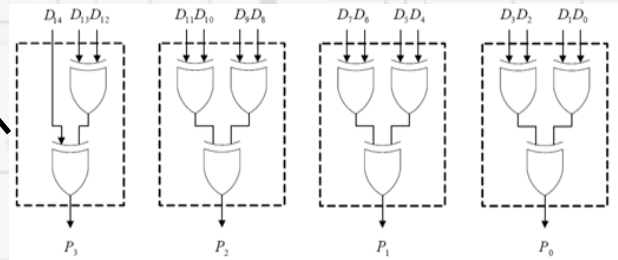


Figure 1.5: 15-bit parameter extractor of the block-XOR PB-CAM.

"Synthesis and Design of Parameter Extractors for Low-Power Pre-computation-Based Content-Addressable Memory Using Gate-Block Selection Algorithm," ASP-DAC, Korea, 21st -24th January, 2008, pp. 316-321

"Synthesis and Design of Parameter Extractors for Low-Power Pre-computation-Based Content-Addressable Memory," IEICE Transactions on Electronics, Vol.E92-C,No.10,pp.1149-1257,Oct. 2009.

Synthesis of Parameter Extractor for Low Power CAM Designs

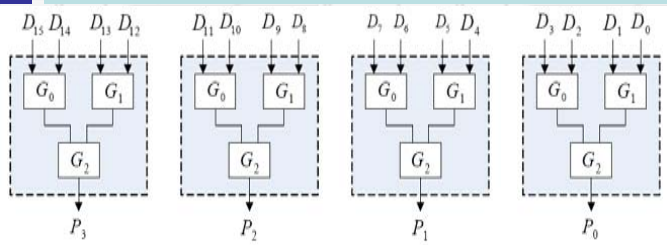
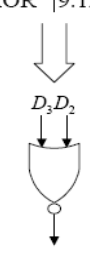


Figure 2.2: 4-bit partition block for 16-bit input data.

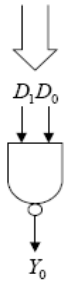
$D_3D_2D_1D_0$	D_3D_2	NAND	NOR	XOR
0 0 1 1	0 0	1	1	0
1 0 0 1	1 0	1	0	1
0 0 0 0	0 0	1	1	0
0 0 1 1	0 0	1	1	0
0 0 1 1	0 0	1	1	0
0 1 0 1	0 1	1	0	1
0 0 1 0	0 0	1	1	0
1 1 1 1	1 1	0	0	0
1 0 0 1	1 0	1	0	1
0 0 1 1	0 0	1	1	0
0 0 1 0	0 0	1	1	0
1 0 0 1	1 0	1	0	1
0 0 0 0	0 0	1	1	0
0 1 1 1	0 1	1	0	1
1 1 1 1	1 1	0	0	0
0 0 1 1	0 0	1	1	0

D_3D_2	C_{avg}
NAND	12.5
NOR	8.125
XOR	9.125



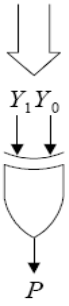
$D_3D_2D_1D_0$	D_1D_0	NAND	NOR	XOR
0 0 1 1	1 1	0	0	0
1 0 0 1	0 1	1	0	1
0 0 0 0	0 0	1	1	0
0 0 1 1	1 1	0	0	0
0 0 1 1	1 1	0	0	0
0 1 0 1	0 1	1	0	1
0 0 1 0	1 0	1	0	1
1 1 1 1	1 1	0	0	0
1 0 0 1	0 1	1	0	1
0 0 1 1	1 1	0	0	0
0 0 1 0	1 0	1	0	1
1 0 0 1	0 1	1	0	1
0 0 0 0	0 0	1	1	0
0 1 1 1	1 1	0	0	0
1 1 1 1	1 1	0	0	0
0 0 1 1	1 1	0	0	0

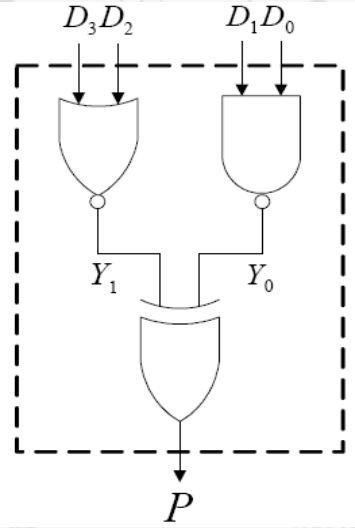
D_1D_0	C_{avg}
NAND	8
NOR	12.5
XOR	8.5



Y_1Y_0	NAND	NOR	XOR
1 0	1	0	1
0 1	1	0	1
1 1	0	0	0
1 0	1	0	1
1 0	1	0	1
0 1	1	0	1
1 1	0	0	0
0 0	1	1	0
0 1	1	0	1
1 0	1	0	1
1 1	0	0	0
0 1	1	0	1
1 1	0	0	0
0 0	1	1	0
0 0	1	1	0
1 0	1	0	1

Y_1Y_0	C_{avg}
NAND	10
NOR	11.125
XOR	8.125



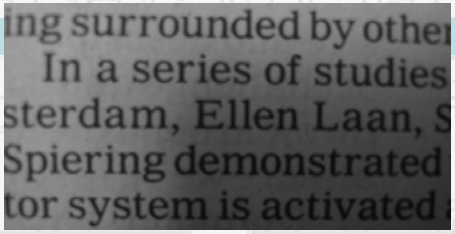


An Efficient Thresholding Algorithm for Document Images Based on Intelligent Block Detection

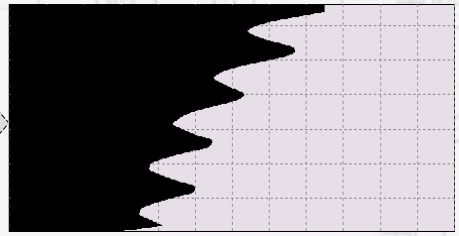
- Image binarization techniques can be divided into two classes: global and local.
  - The global methods use one calculated threshold value to divided image pixels into object or background class.
    - Otsu's algorithm [1978]
      - Fast but poor result
  - The local approaches compute a separate threshold for each pixel based on its neighbors.
    - Niblack's algorithm [1986]
    - Sauvola's algorithm [2000]
    - Gatos' algorithm [2006]
      - Slow but high quality

"An Efficient Thresholding Algorithm for Degraded Document Images based on Intelligent Block Detection," IEEE SMC 2008, Singapore, 12 - 15 October 2008, pp. 667-672
"Adaptive Thresholding Algorithm: Efficient Computation Technique Based on Intelligent Block Detection for Degraded Document Images" Pattern Recognition, Vol. 43, Issue 9, September 2010, Pages 3177-3187

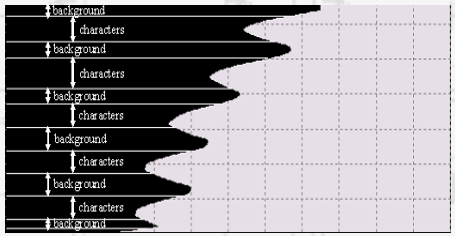
An Efficient Thresholding Algorithm for Document Images Based on Intelligent Block Detection



Gray-level image



Histogram analysis



Block size detection

Block Detect

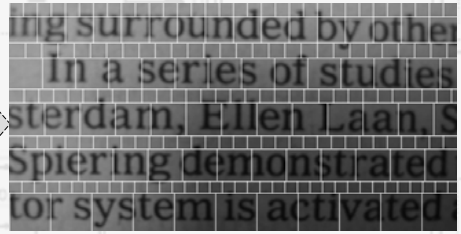
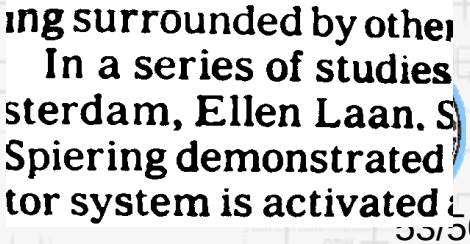


Image segmentation

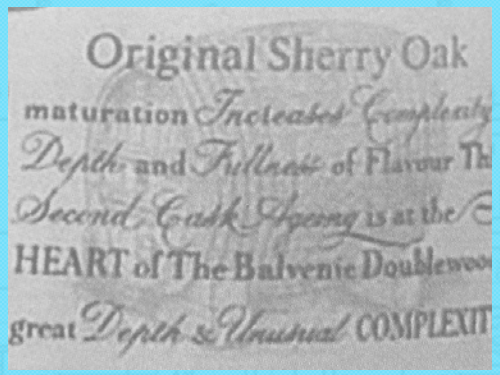
Image thresholding

Binarization

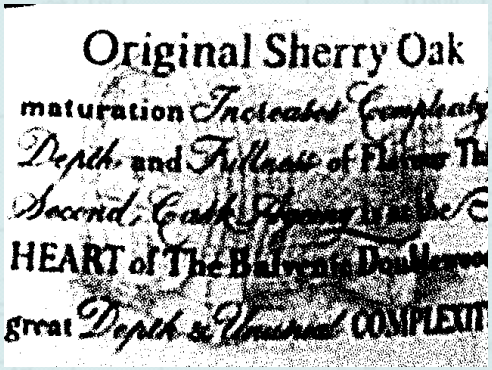
Binary document image



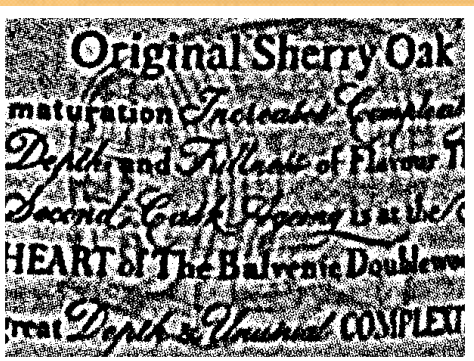
An Efficient Thresholding Algorithm for Document Images Based on Intelligent Block Detection



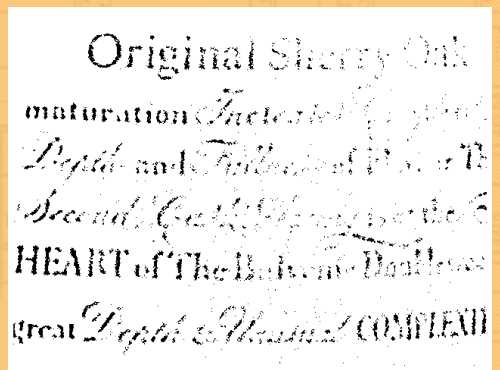
Original image



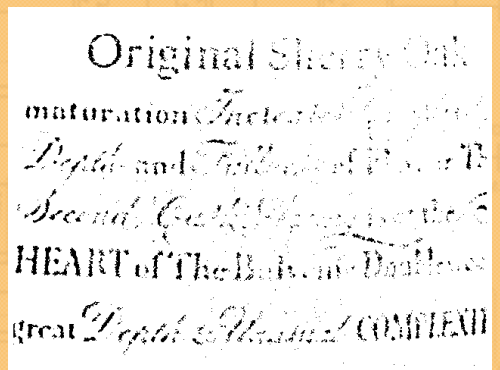
Otsu's method



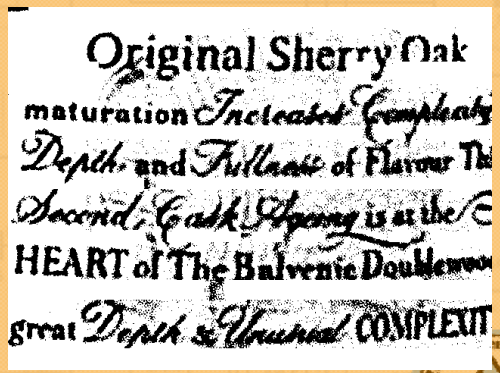
Niblack's method



Sauvola's method

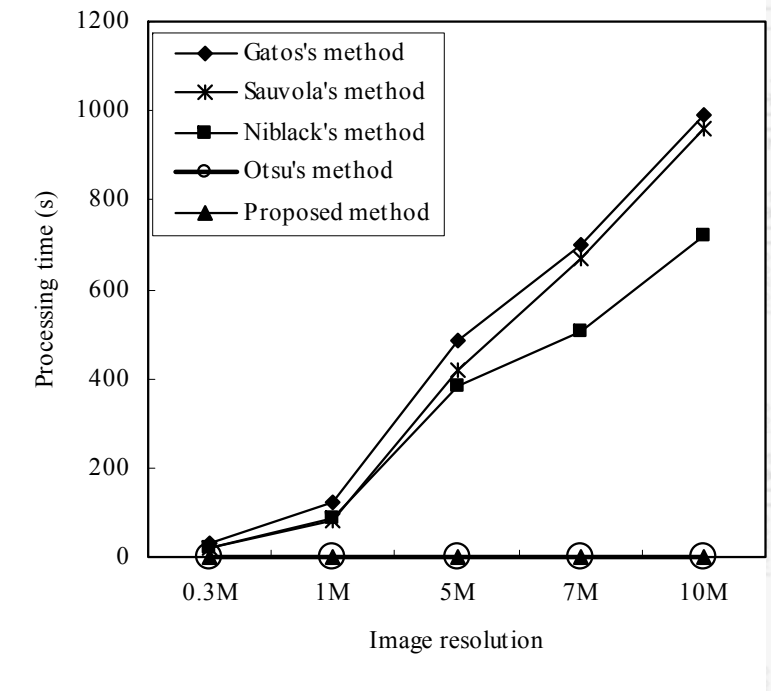


Gatos' method

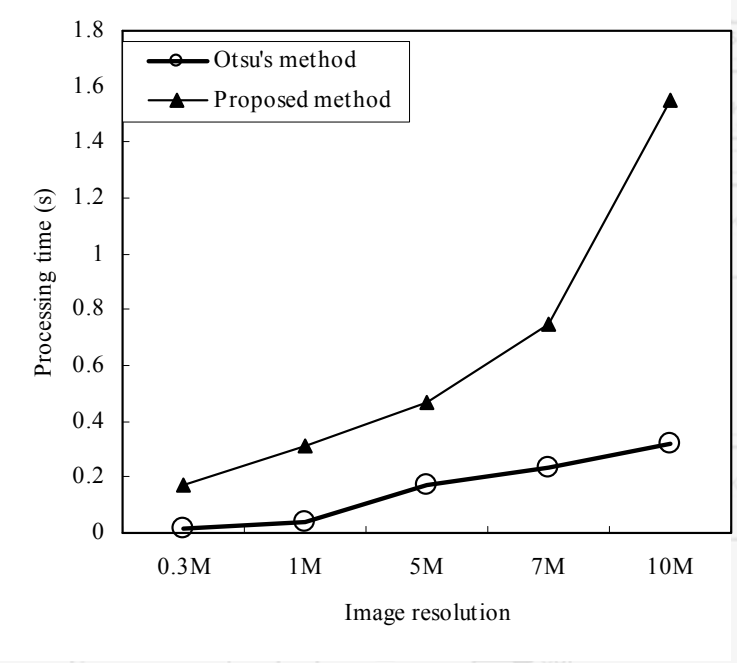


Proposed method

An Efficient Thresholding Algorithm for Document Images Based on Intelligent Block Detection



The tendency of processing time upon various image resolution among Gatos', Sauvola's, Niblack's, Otsu's, and the proposed methods.



The tendency of processing time upon various image resolution between Otsu's and the proposed methods.

Thank You for Your Attention~

TAIWAN TECH