



Introduction to Digital Signal Processing Systems

Shao-Yi Chien

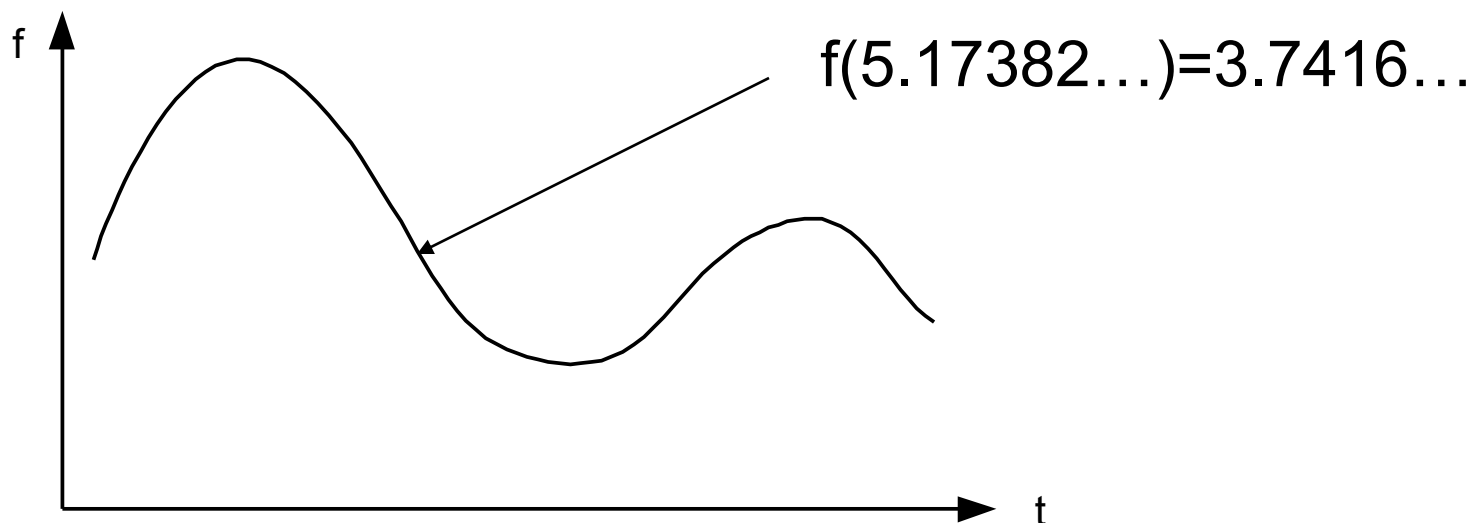


Outline

- Introduction
- Typical DSP algorithms
- Scaled CMOS technologies
- Representations of DSP algorithms

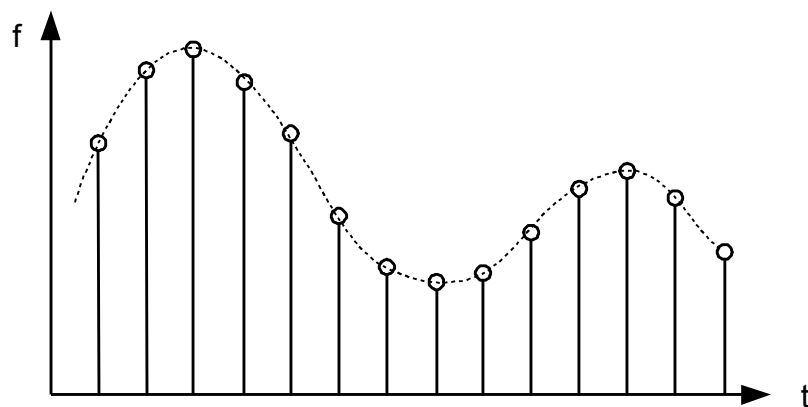
Analog Signal

- Real-world signal
- Infinite accuracy on time and magnitude

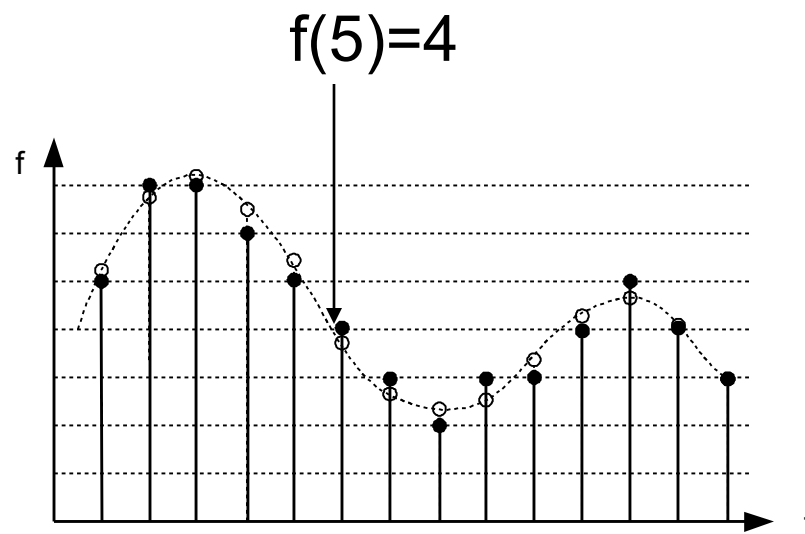


Digital Signal

- Get after sampling and quantization
- Finite accuracy on time and magnitude
- Easy to process with digital processing element

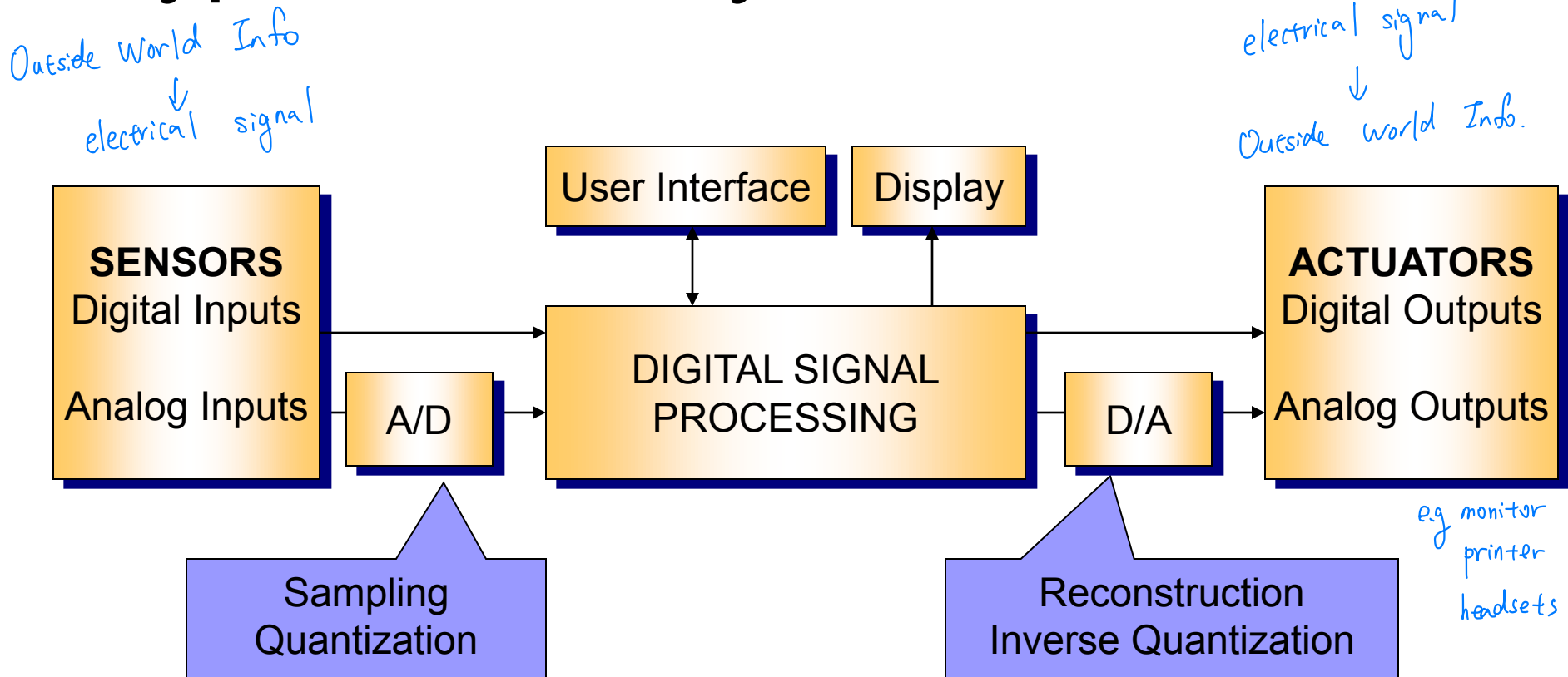


Sampling



Quantization

Typical DSP Systems





Advantages of Analog Signal Processing

- Can operate in very high frequency
- Sometimes low area *e.g.) Amplification requires dedicated amplification circuit.*
- Low power



Advantages of Digital Signal Processing (DSP)

- More robust
 - Insensitive to environment and component tolerance
- The accuracy can be controlled better
- Can cancel the noise and interference while amplifying the signal
- Predictable, repeatable behavior
 - Can be stored and recovered, transmitted and received, processed and manipulated without error



Features of DSP Systems

- Real-time throughput requirement
 - So-called hard real-time systems
- Data-driven property *Fire the calculation when the data arrives.*
- Non-terminating program *analogous to a forever loop.*

Hard Real-Time Systems





Performance Metrics of DSP Systems

- PPA {
- Hardware circuitry and resources (area)
 - Speed of execution
 - Power consumption
 - Finite word length performance

eg. floating point word size, accuracy.

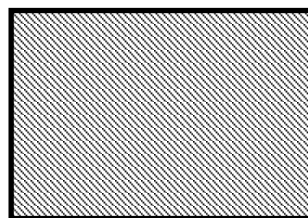
Characteristics of DSP Systems (1/4)

■ Data format

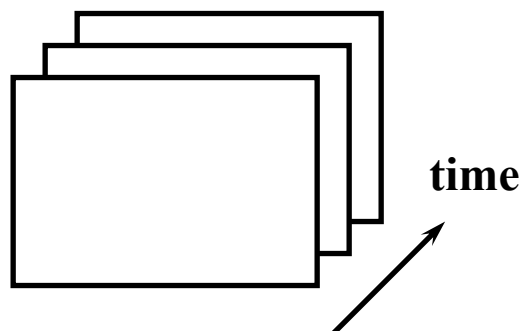
□ 1D speech



□ 2D image



□ 3D video



Characteristics of DSP Systems

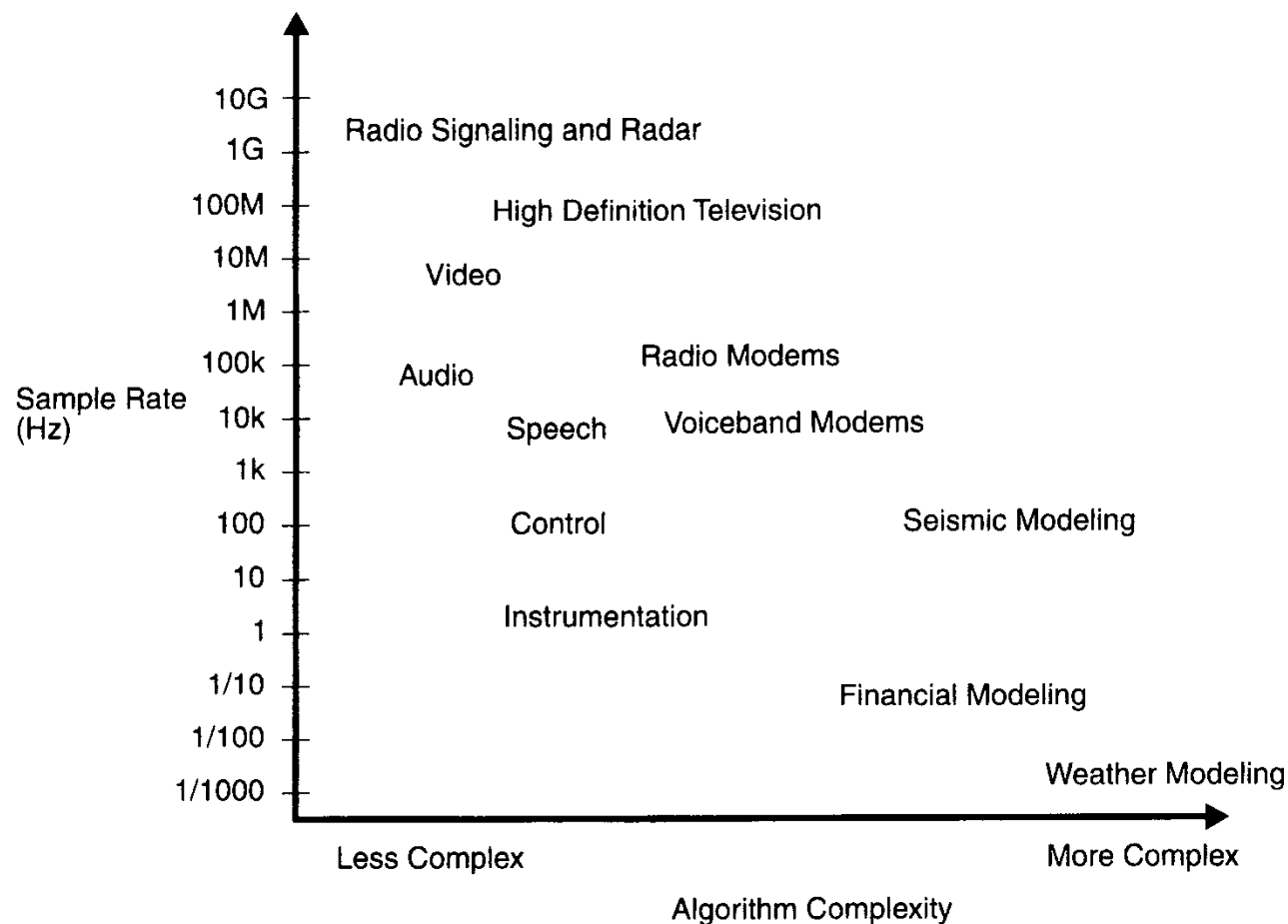
(2/4)

■ Algorithms

DSP Algorithm	System Application
Speech coding and decoding	Digital cellular telephones, personal communications systems, digital cordless telephones, multimedia computers, secure communications
Speech encryption and decryption	Digital cellular telephones, personal communications systems, digital cordless telephones, secure communications
Speech recognition	Advanced user interfaces, multimedia workstations, robotics, automotive applications, digital cellular telephones, personal communications systems, digital cordless telephones
Speech synthesis	Multimedia PCs, advanced user interfaces, robotics
Speaker identification	Security, multimedia workstations, advanced user interfaces
Hi-fi audio encoding and decoding	Consumer audio, consumer video, digital audio broadcast, professional audio, multimedia computers
Modem algorithms	Digital cellular telephones, personal communications systems, digital cordless telephones, digital audio broadcast, digital signaling on cable TV, multimedia computers, wireless computing, navigation, data/facsimile modems, secure communications
Noise cancellation	Professional audio, advanced vehicular audio, industrial applications
Audio equalization	Consumer audio, professional audio, advanced vehicular audio, music
Ambient acoustics emulation	Consumer audio, professional audio, advanced vehicular audio, music
Audio mixing and editing	Professional audio, music, multimedia computers
Sound synthesis	Professional audio, music, multimedia computers, advanced user interfaces
Vision	Security, multimedia computers, advanced user interfaces, instrumentation, robotics, navigation
Image compression and decompression	Digital photography, digital video, multimedia computers, video-over-voice, consumer video
Image compositing	Multimedia computers, consumer video, advanced user interfaces, navigation
Beamforming	Navigation, medial imaging, radar/sonar, signals intelligence
Echo cancellation	Speakerphones, modems, telephone switches
Spectral estimation	Signals intelligence, radar/sonar, professional audio, music

Characteristics of DSP Systems (3/4)

■ Sample rates





Characteristics of DSP Systems (4/4)

- Clock rates
- Numeric representations
 - Two's complements
 - One's complements
 - Gray Code
 - ...



Standard Digital Signal Processors (1/2)

- Allow rapid prototyping and time-to-market
- Sometimes, the execution speed and code size is reasonably good
- Not always cost effective
- Often cannot meet the requirements of throughput, power consumption, and size

Standard Digital Signal Processors (2/2)

■ DSP Architectures

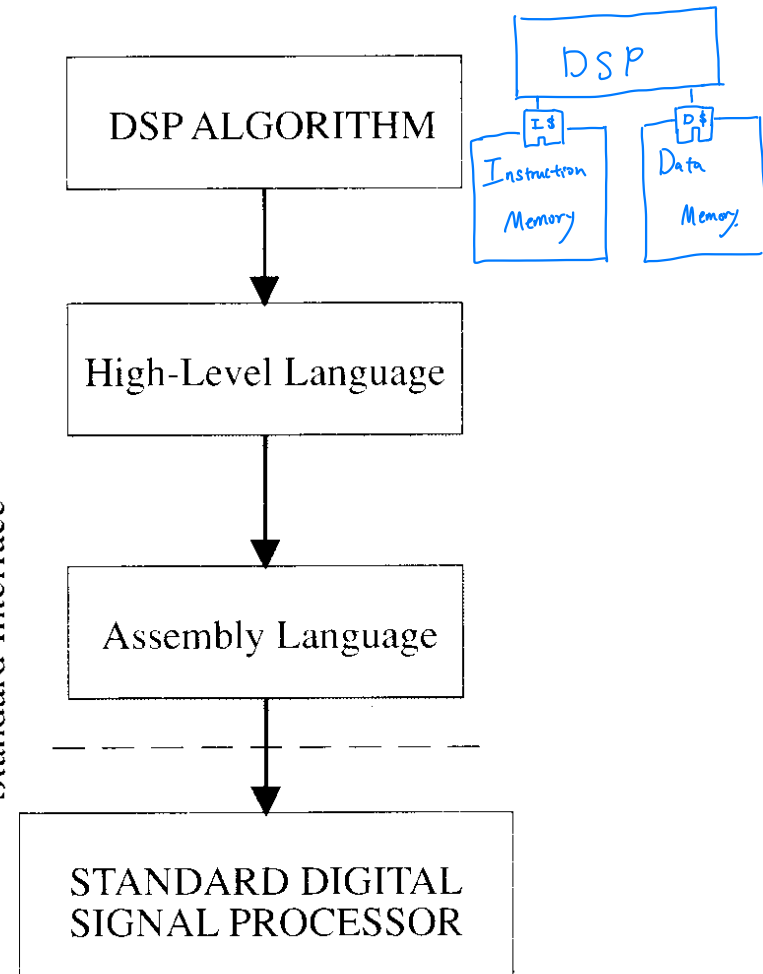
- Harvard architecture
- MAC *Multiply - And - Accumulate*
- Fixed-point arithmetic
more efficient than floating-point

■ Alternatives:

- DSP-enhanced CPU
- GPU *opt. for Matrix calculation*

*Specialised Instruction
(e.g. fmadd) to
initiate DSP. in SoC*

Standard Interface



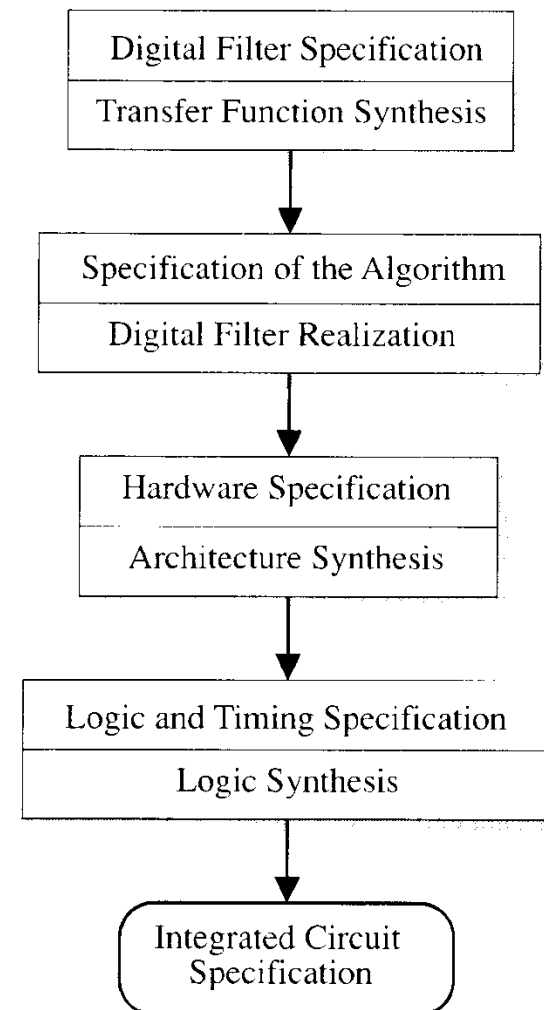


Application-Specific ICs for DSP (1/2)

- Better performances
 - ☐ Processing capacity
 - ☐ Power consumption
 - ☐ Pin-restriction problem
- Main problem: the system is very complex to design
 - ☐ Long time-to-market

Application-Specific ICs for DSP (2/2)

- Large design space
- Hard to find optimal solution
- System → specification → algorithm → hardware architecture → logic implementation → VLSI implementation
- ASIC → Accelerator design in an SoC



Typical DSP Algorithms

- Convolution
- Correlation
- Digital filters
- Adaptive filters
- Motion estimation
- Discrete cosine transform (DCT)
- Vector quantization (VQ)
- Viterbi algorithm and dynamic programming
- Decimator and expander
- Wavelets and filter banks

Convolution (1/2)

$$y(n) = x(n) * h(n) = \sum_{k=-\infty}^{\infty} x(k)h(n - k).$$

- Can be used to describe the behavior of a linear time-invariant systems
 - $x(n)$: input signal
 - $y(n)$: output signal
 - $h(n)$: unit-sample response

Convolution (2/2)

- Finite impulse response (FIR) system

$$h(n) = \frac{1}{M_1 + M_2 + 1} \sum_{-M_1}^{M_2} \delta(n - k)$$

- Infinite impulse response (IIR) system

$$h(n) = \sum_{k=-\infty}^n \delta(k)$$

Digital Filters

- LTI, causal filter

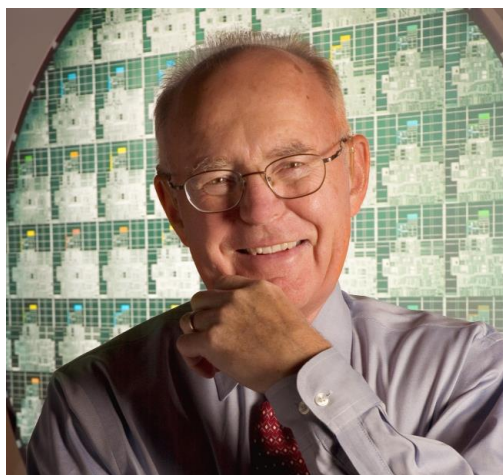
$$y(n) = - \sum_{k=1}^N a_k y(n-k) + \sum_{k=0}^{M-1} b_k x(n-k).$$

- M-tap finite impulse response filter

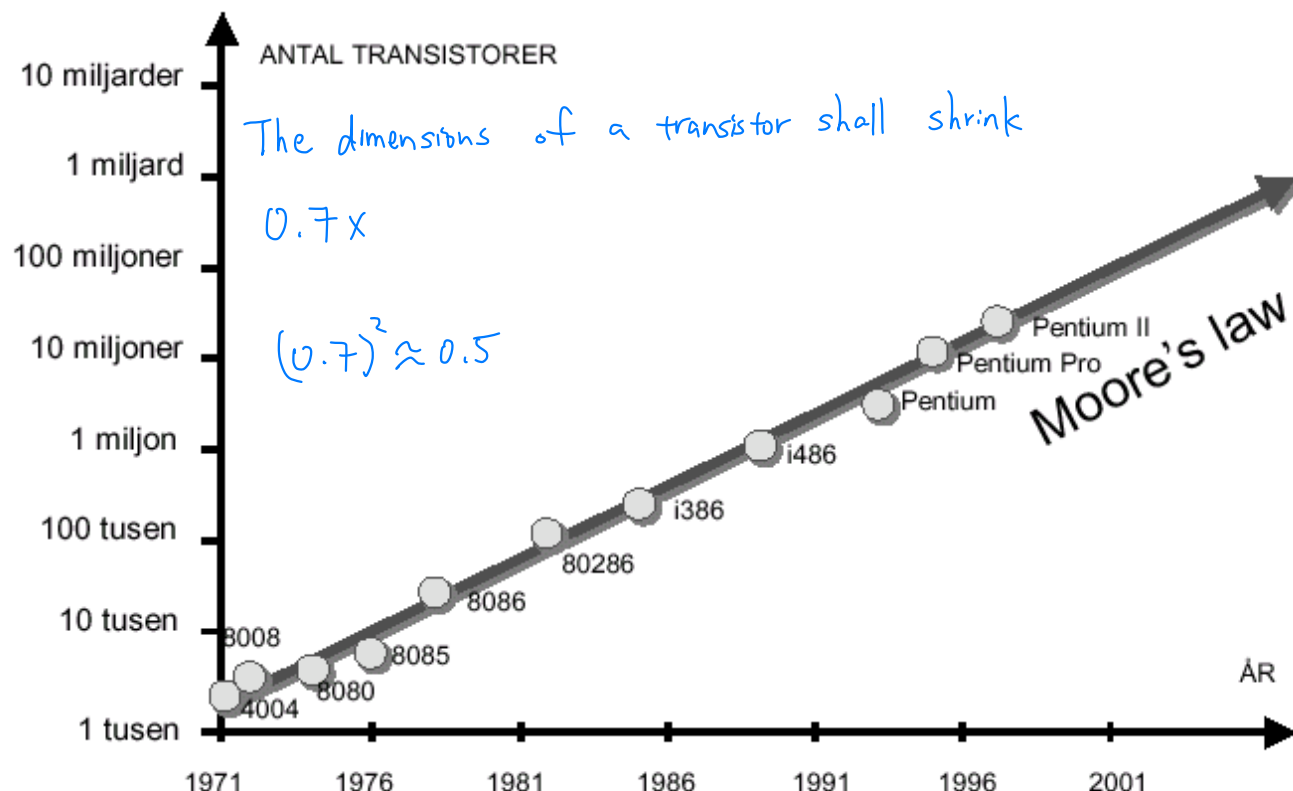
$$y(n) = \sum_{k=0}^{M-1} b_k x(n-k)$$

Chip Development

Moore's law:
The number of transistors per chip doubles every 18 months.



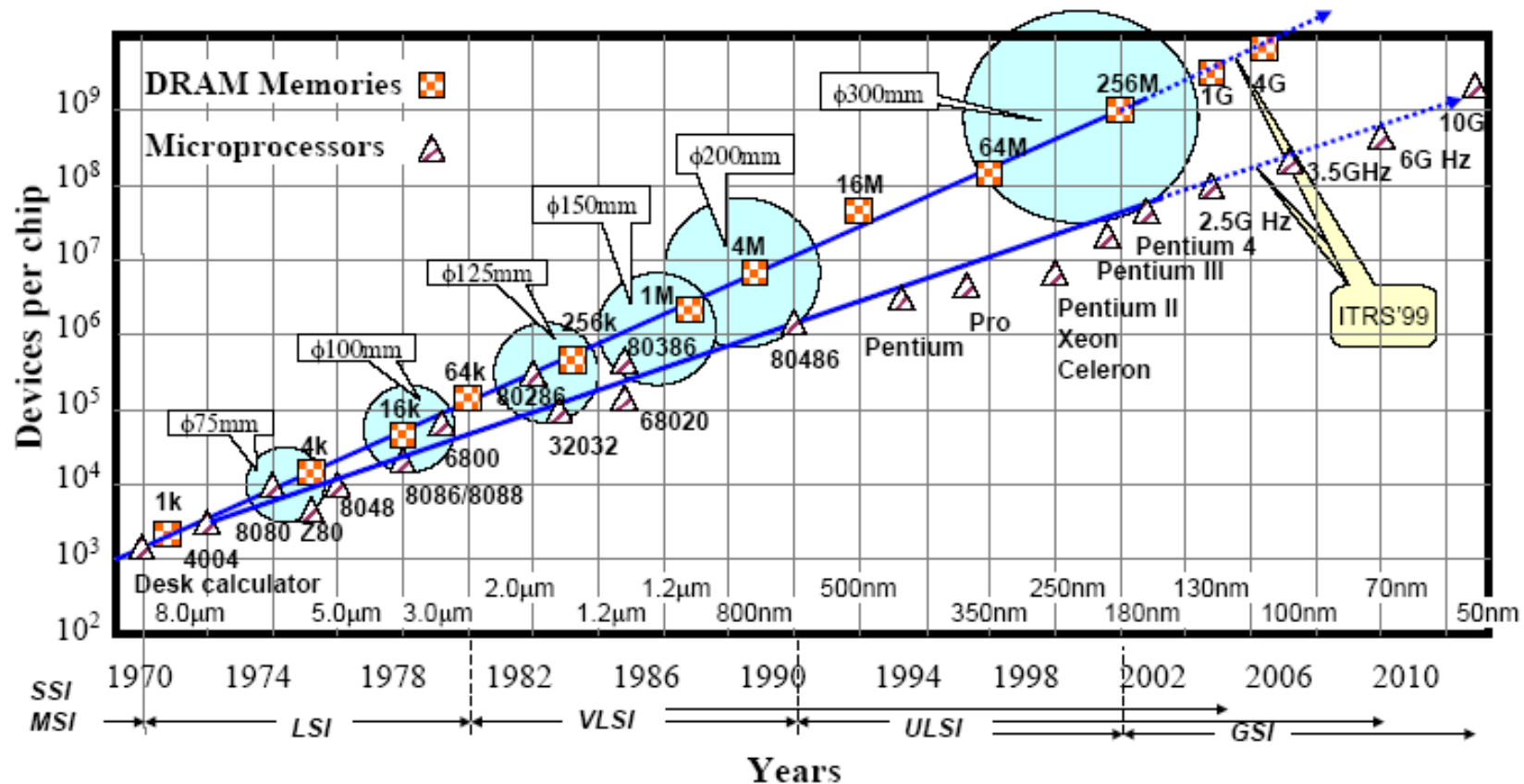
Gordon Moore
One of the founders of Intel



Technology roadmap:

<http://notes.sematech.org/ntrs/PubINTRS.nsf>

Moore's Law



Scaled CMOS technology (Moore's Law) (1/2)

Year of Production:	2001	2003	2005	2007	2010	2016
DRAM Half-Pitch [nm]:	130	100	80	65	45	22
Overlay Accuracy [nm]:	46	35	28	23	18	9
MPU Gate Length [nm]:	90	65	45	35	25	13
CD Control [nm]:	8	5.5	3.9	3.1	2.2	1.1
T _{OX} (equivalent) [nm]:	1.3-1.6	1.1-1.6	0.8-1.3	0.6-1.1	0.5-0.8	0.4-0.5
Junction Depth [nm]:	48-95	33-66	24-47	18-37	13-26	7-13
Metal Cladding [nm]:	16	12	9	7	5	2.5
Inter-Metal Dielectric K:	3.0-3.6	3.0-3.6	2.6-3.1	2.3-2.7	2.1	1.8



Scaled CMOS technology (Moore's Law) (2/2)

2022 IRDS ORTC							
YEAR OF PRODUCTION	2021	2022	2025	2028	2031	2034	2037
Logic device technology naming note definition [1a]	G51M29	G48M24	G45M20	G42M16	G40M16T2	G38M16T4	G38M16T6
Logic industry "Node Range" Labeling (nm) [2]	"5"	"3"	"2"	"1.5"	"1.0-eq"	"0.7nm-eq"	"0.5nm-eq"
Fine-pitch 3D integration scheme		Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Platform device for logic [1b]	FinFET	FinFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
LOGIC CELL AND FUNCTIONAL FABRIC TARGETS							
Digital block area scaling	1.00	1.00	0.74	0.55	0.26	0.13	0.08
LOGIC DEVICE GROUND RULES							
MPU/SoC M0 1/2 Pitch (nm) [3]	15	12	10	8	8	8	8
Gate length (nm) [4]	17	16	14	12	12	12	12
Lateral GAA (nanosheet) Minimum Thickness (nm)		1	3	3	4	4	4
Number of stacked tiers [5]		1	1	1	2	4	6
Number of stacked nanosheets in one device [5]		1	3	3	4	4	4
LOGIC DEVICE Electrical							
Vdd (V) [6]	0.75	0.70	0.65	0.65	0.60	0.60	0.60
DRAM TECHNOLOGY							
DRAM Min half pitch (nm) [7]	17.5	15.5	13	14	11.5	10	10
DRAM cell size (μm^2) [8]	0.00184	0.00165	0.00118	0.00085	0.00062	0.00044	0.00025
DRAM storage node cell capacitor voltage (V) [9]	0.50	0.45	0.45	0.43	0.4	0.4	0.4
NAND Flash							
Flash 2D NAND Flash uncontacted poly 1/2 pitch – F (nm) 2D [10a,b]	15	15	15	15	15	15	15
Product highest density (3D) (commercialized) [11]	1T	1.3T	2.6T	4T	6T	8T	12T
Flash Product Maximum bits/cell (2D_3D) [12]	2_4	4	5	5	6	6	6
Flash 3D NAND Maximum Number of Memory Layers [13]	64-97	128-192	256-384	384-576	576-768	768-1024	1024-1536
Maximum chip size (mm^2) [14]	140	140	140	140	140	140	140

[International Roadmap for Devices and Systems (IRDS) 2022]



DSP and VLSI

■ Modern DSP







- Well suite to VLSI implementation
- Feasible or economically viable only if implemented using VLSI technologies

■ VLSI

- Large investment → need large volume of products
 - Communication
 - Consumer applications
- Necessary performance requirement (especially real-time requirement)
 - DSP systems are hard real-time systems

Demands for Semiconductors

2019 DEMAND BY END-USE

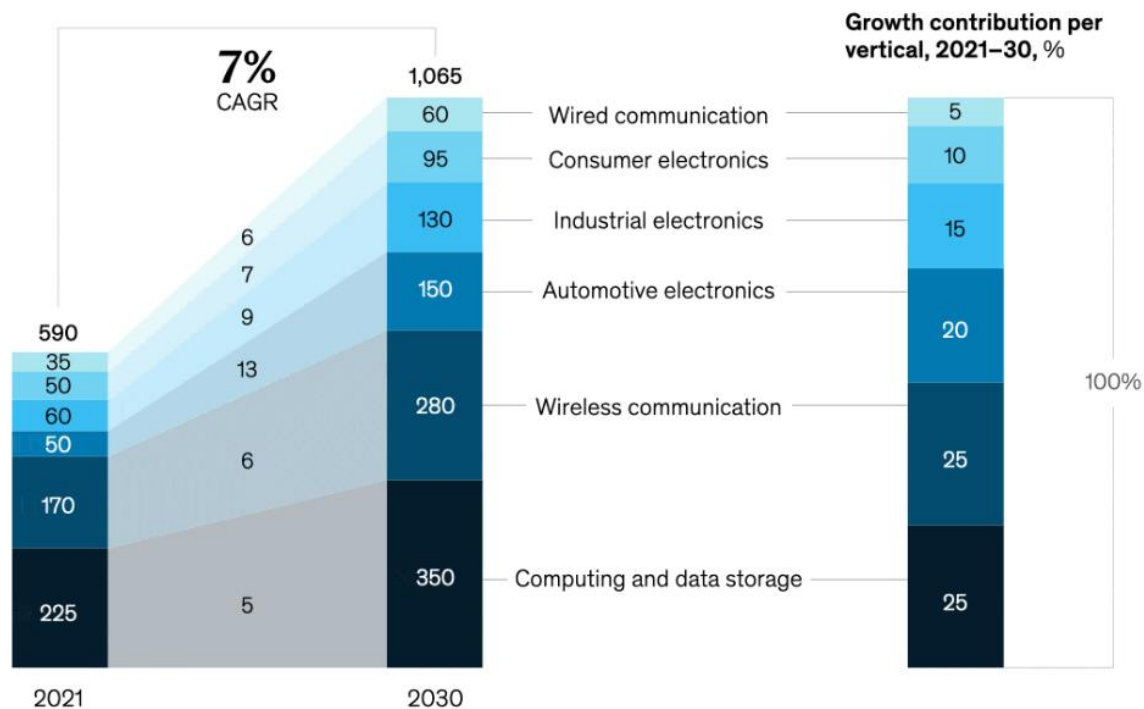
End-Use Category						
	Communication	Computer	Consumer	Automotive	Industrial	Government
Annual Growth	-10.5	-18.7	-5.2	-6.9	-13.0	13.0
Total Value (\$B)	136.0	117.3	54.7	50.2	48.9	5.2

[STATE OF THE U.S. SEMICONDUCTOR INDUSTRY 2020]

Demands for Semiconductors

The overall growth in the global semiconductor market is driven by the automotive, data storage, and wireless industries.

Global semiconductor market value by vertical, indicative, \$ billion



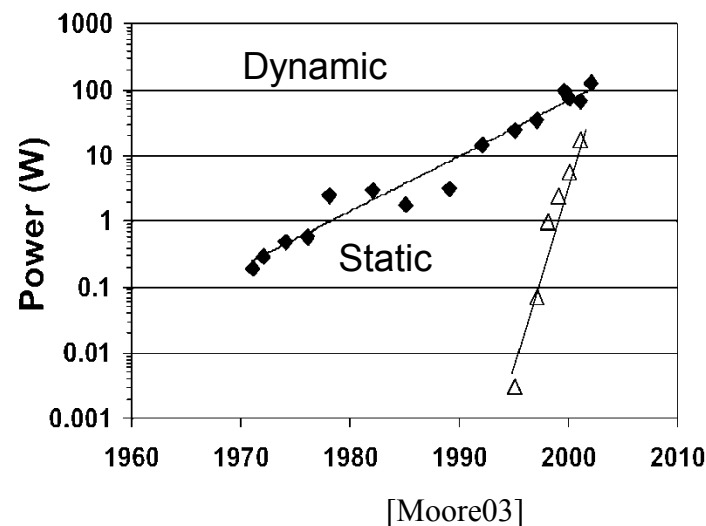
Note: Figures are approximate.

McKinsey
& Company

[Mckinsey 2022]

Problems: Increasing Static Power

- V_{DD} decreases
 - Save dynamic power
 - Protect thin gate oxides and short channels
 - No point in high value because of velocity sat.
- V_t must decrease to maintain device performance
- But this causes exponential increase in OFF leakage
- Major future challenge



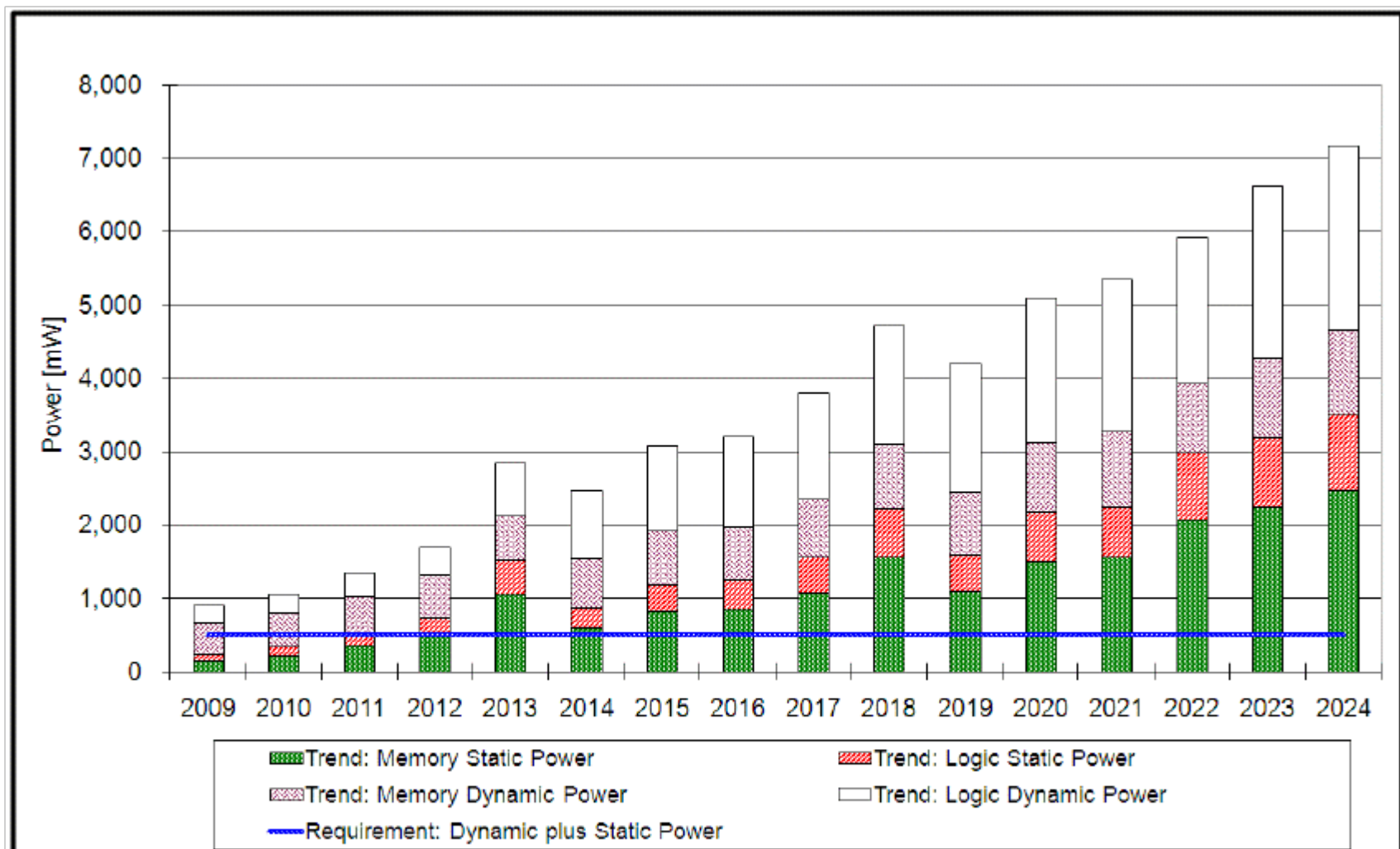
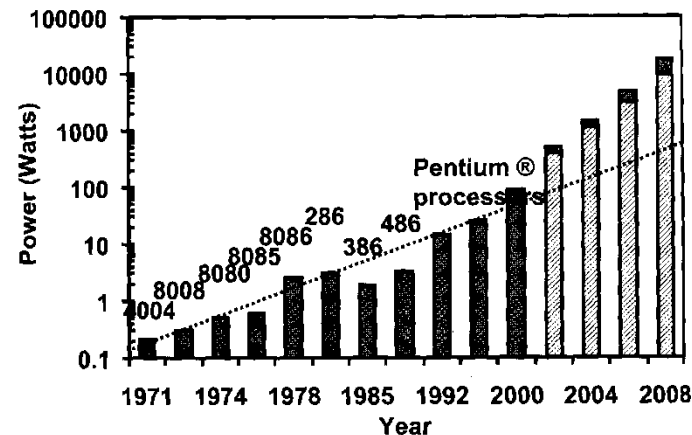


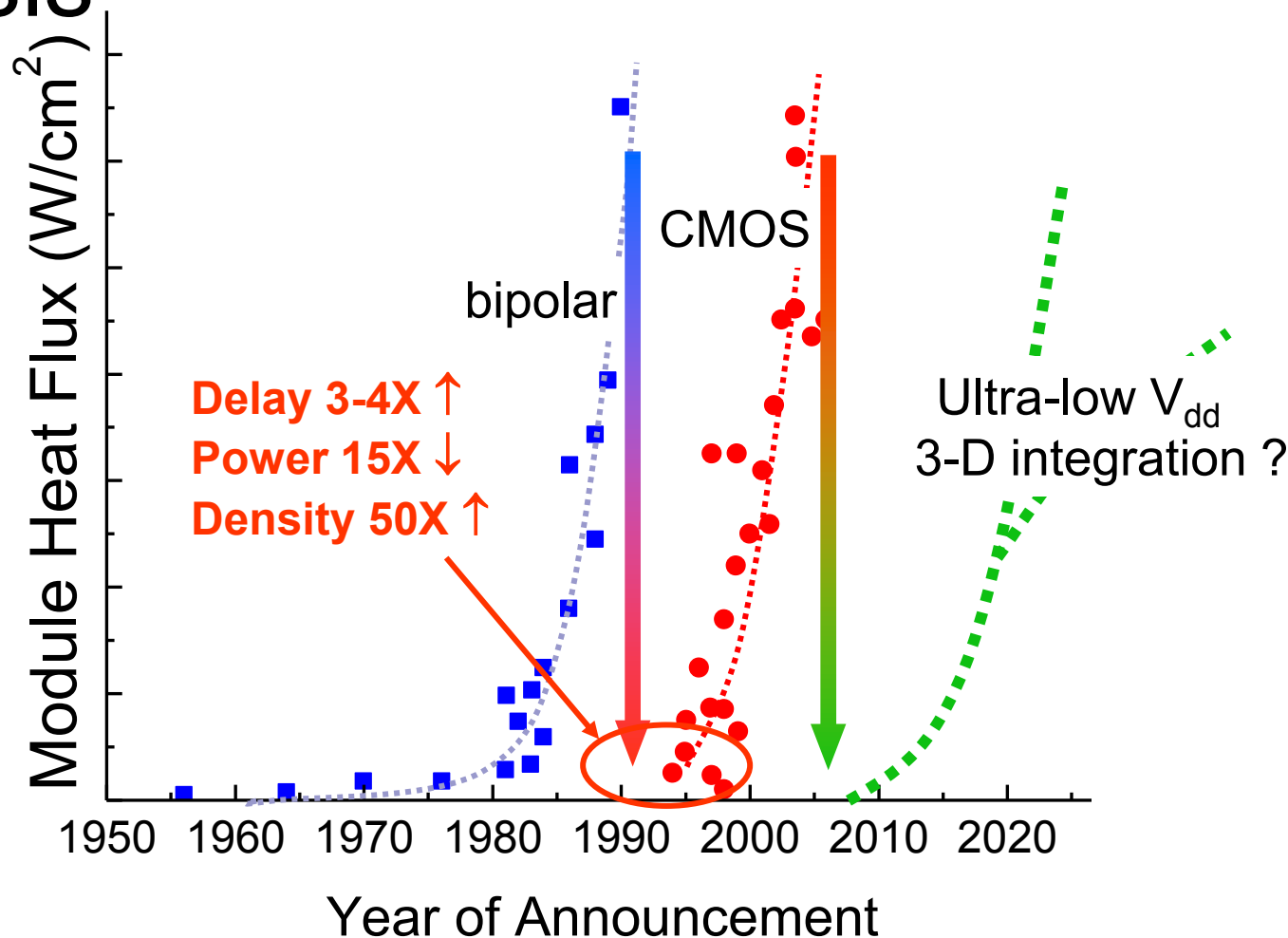
Figure SYSD6 SOC Consumer Portable Power Consumption Trends—UPDATED

Problems: Power Density

- 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

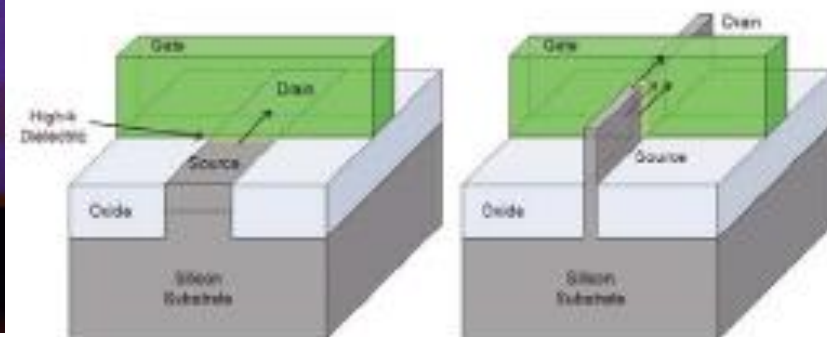


Problems: Scaling and the Power Crisis



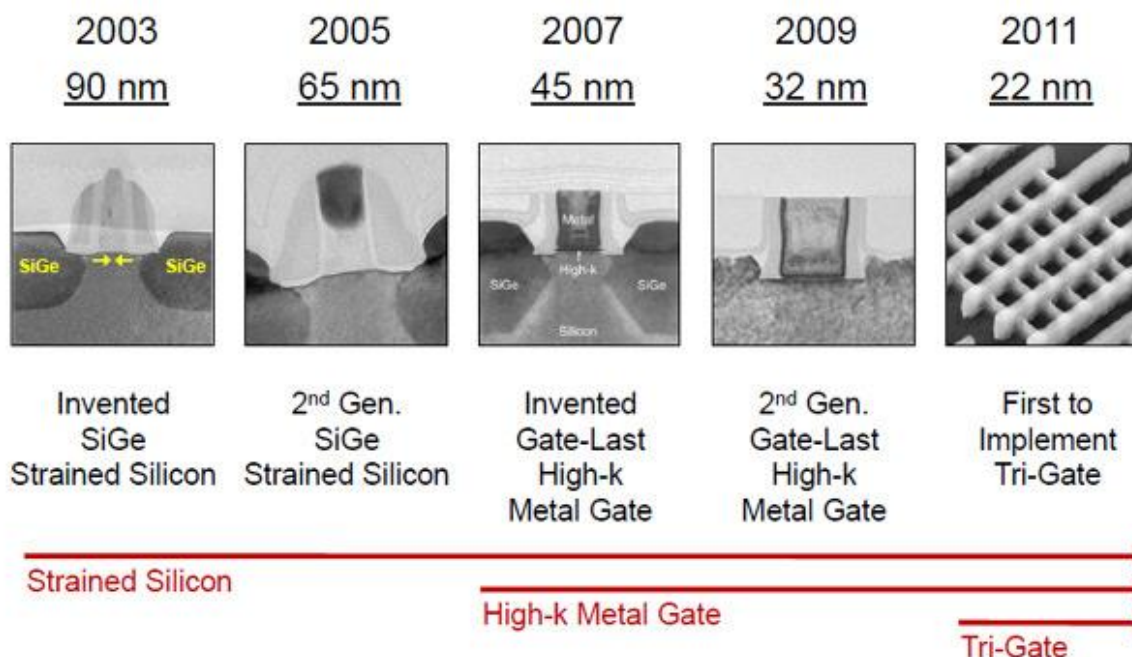
After: R. Schmidt *et al.*, IBM J. R&D, (2002).

FinFET



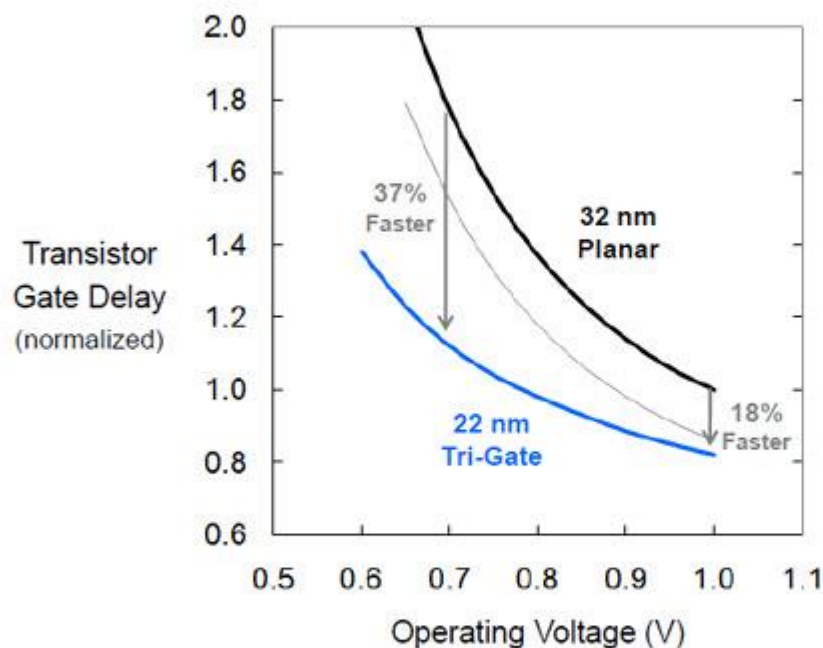
FinFET

Intel Transistor Leadership



FinFET

Transistor Gate Delay



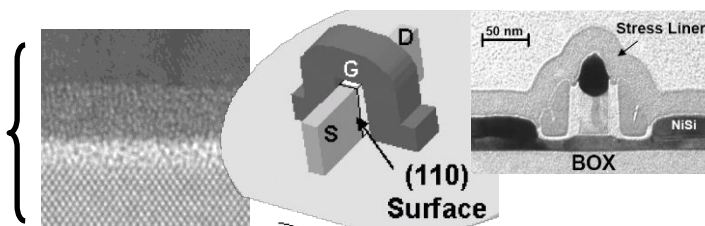
22 nm Tri-Gate transistors provide improved performance at high voltage
and an *unprecedented* performance gain at low voltage



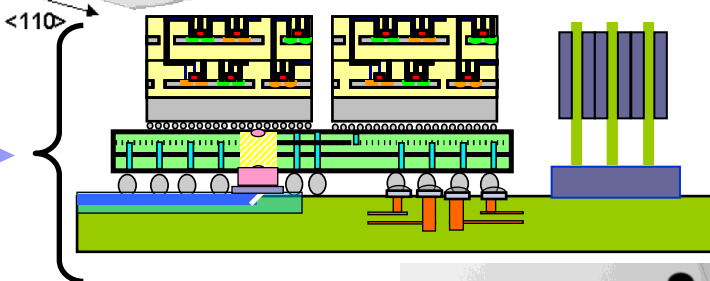
Science and Technology Strategy / Roadmap

2000 2005 2010 2015 2020 2025 2030

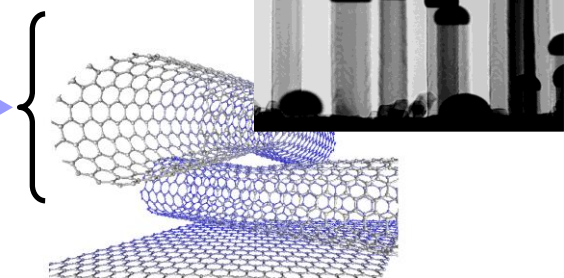
Plan A: Extending Si CMOS



Plan B: Subsystem Integration



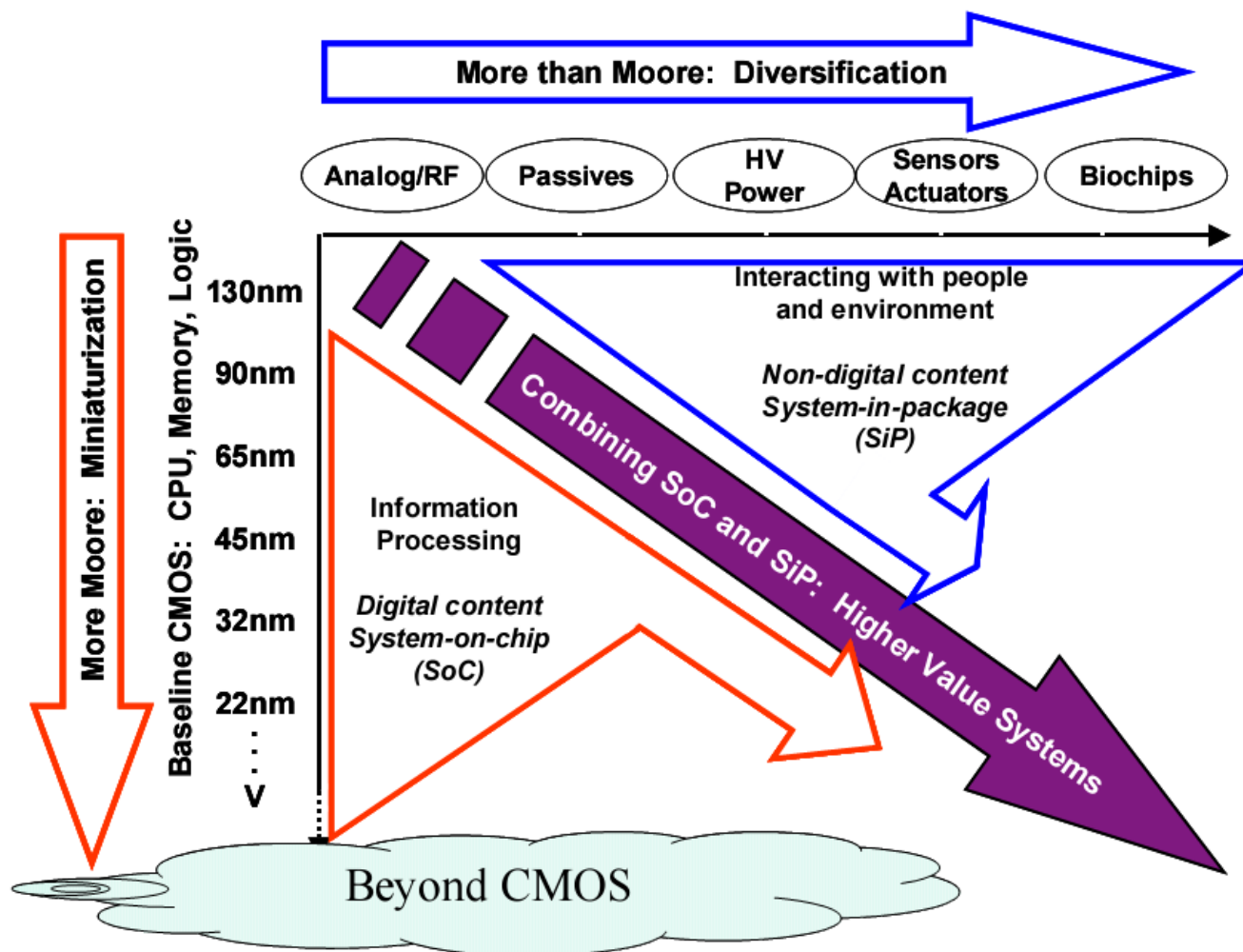
Plan C: Post Si CMOS Options



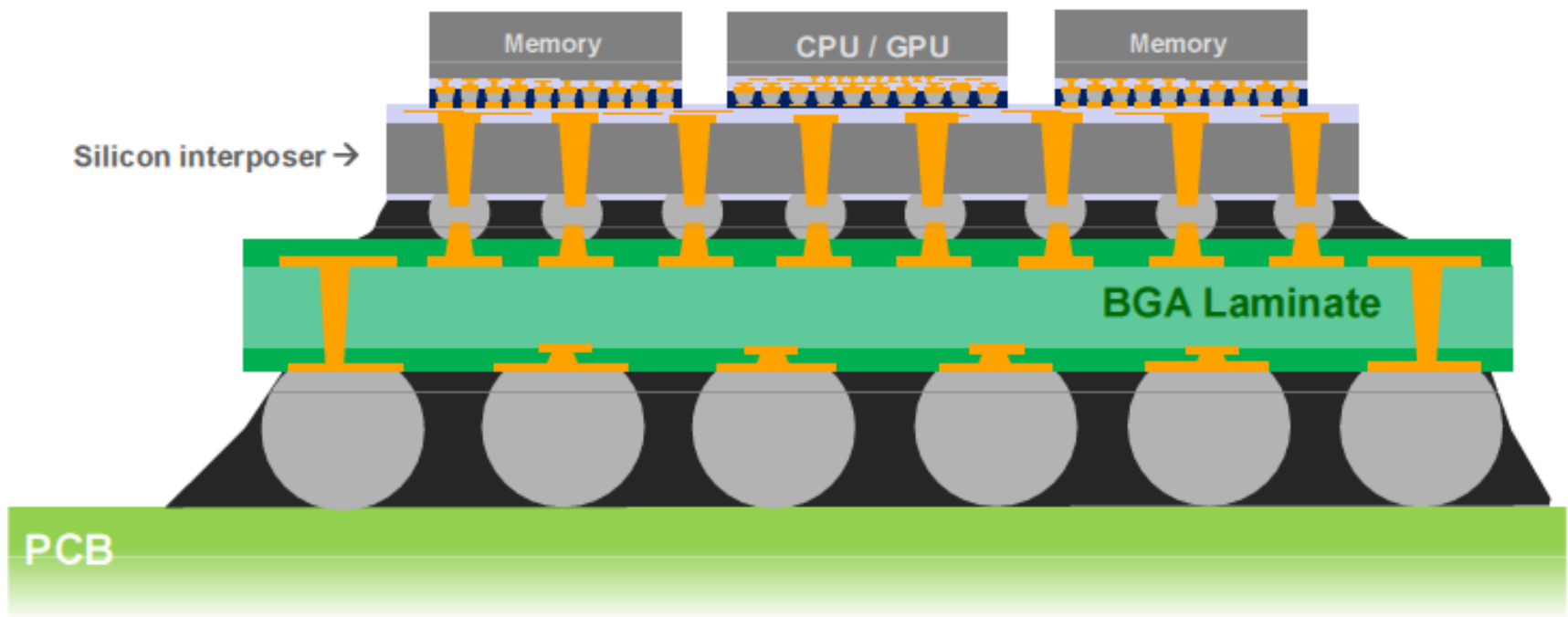
Plan Q: Quantum Computing



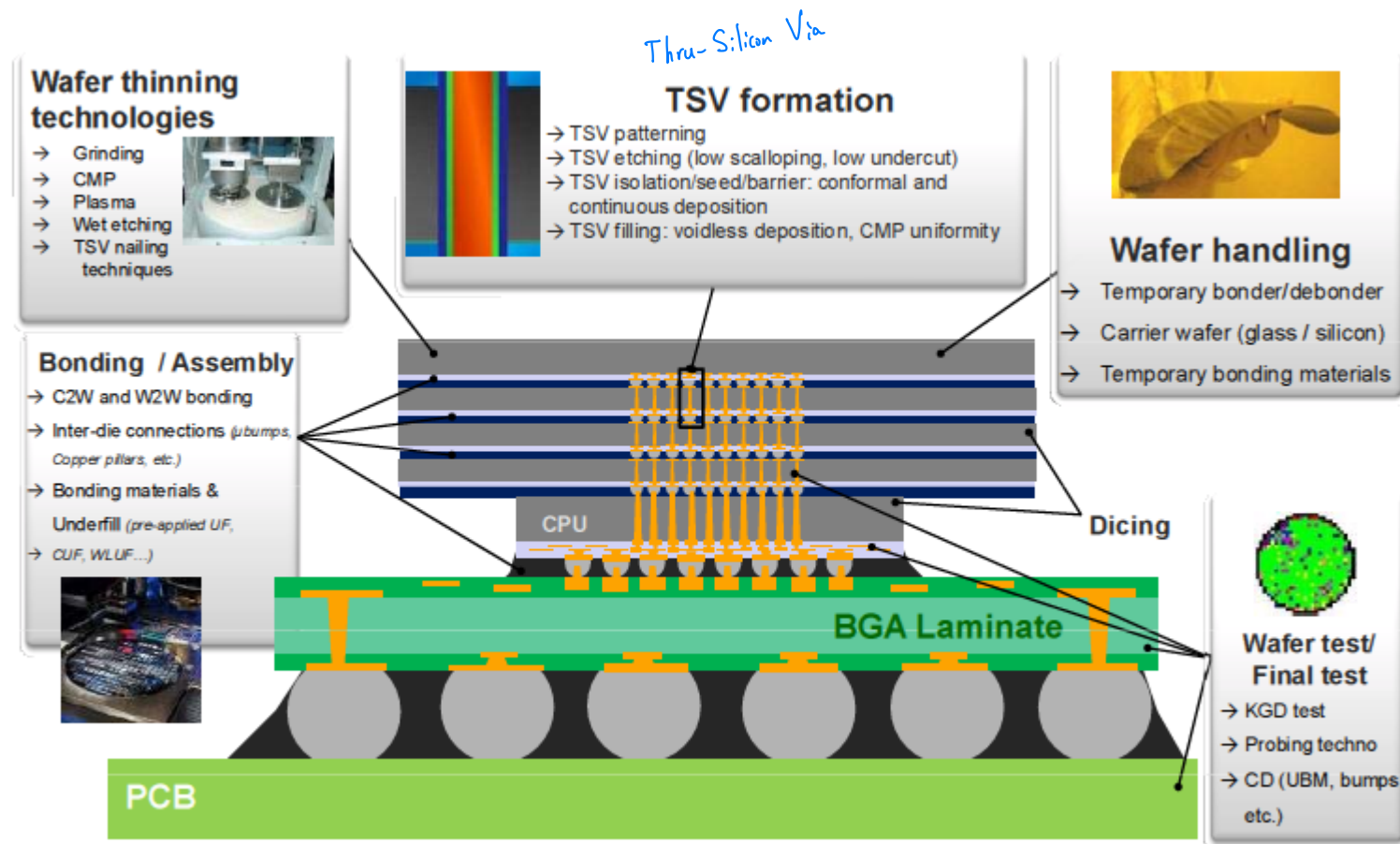
More Moore & More than Moore !!!



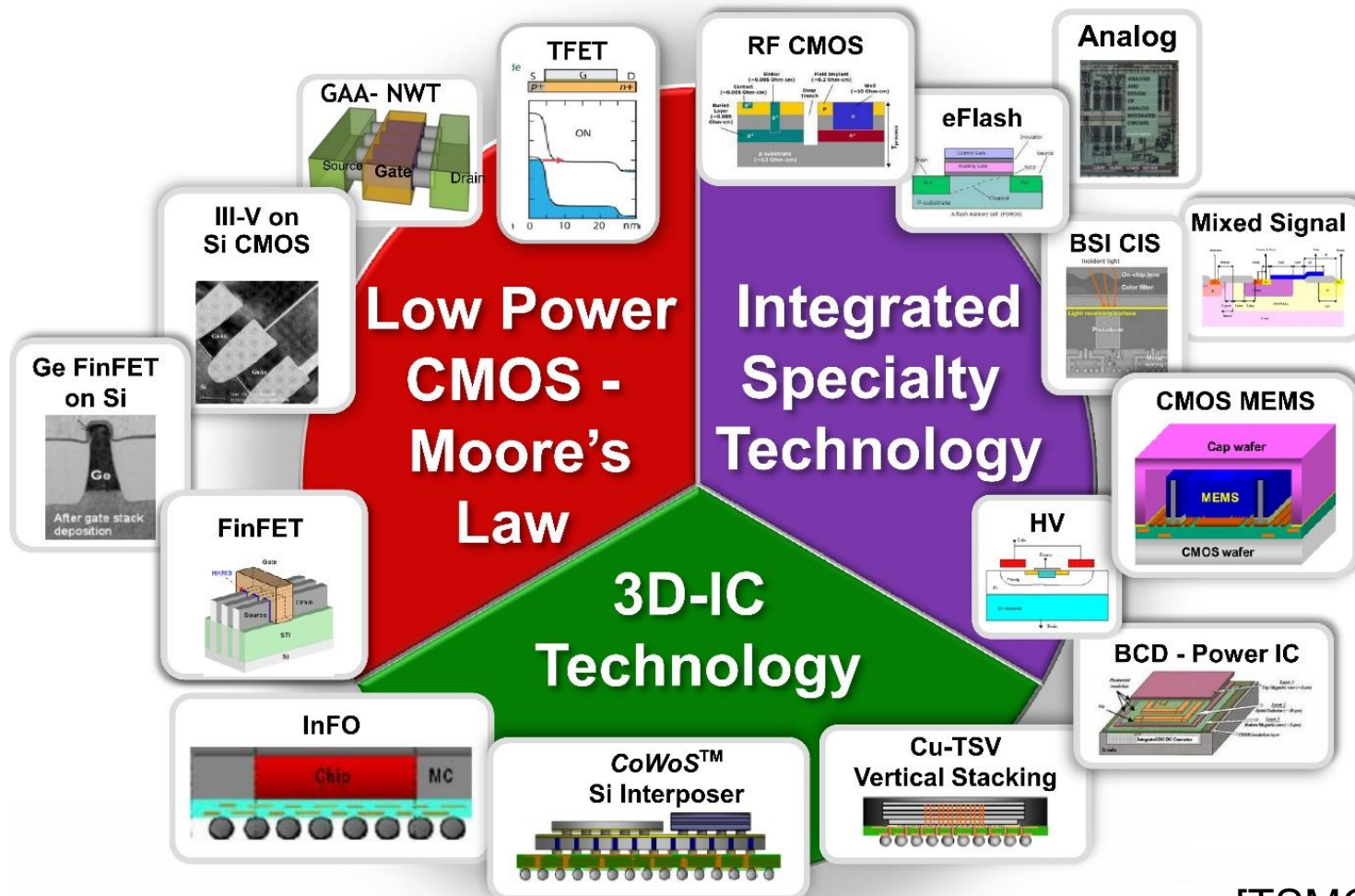
2.5D Interposer



3D-IC Technology



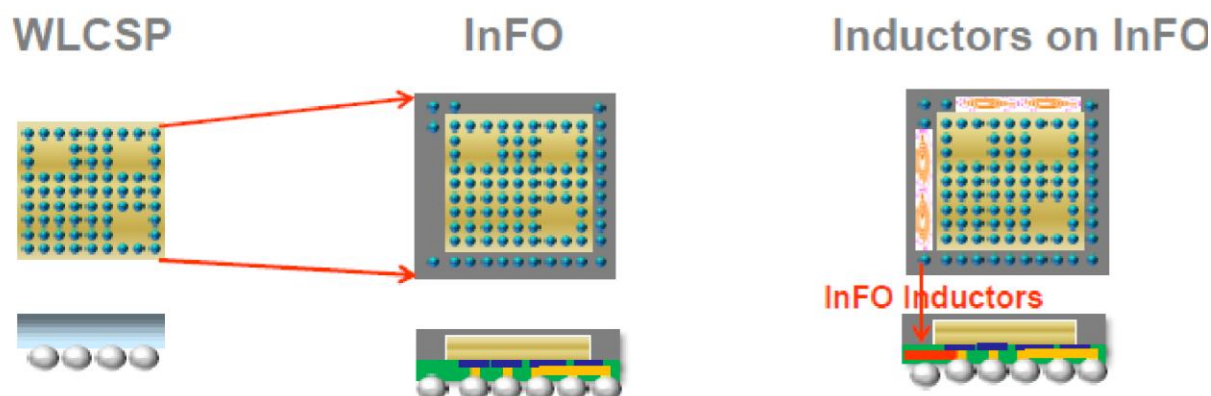
Heterogeneous System Integration



[TSMC 2015]

InFO (Integrated Fan Out)

- **Single-Die InFO:** An extension of “Fan-in” WLCSP to enable more IOs



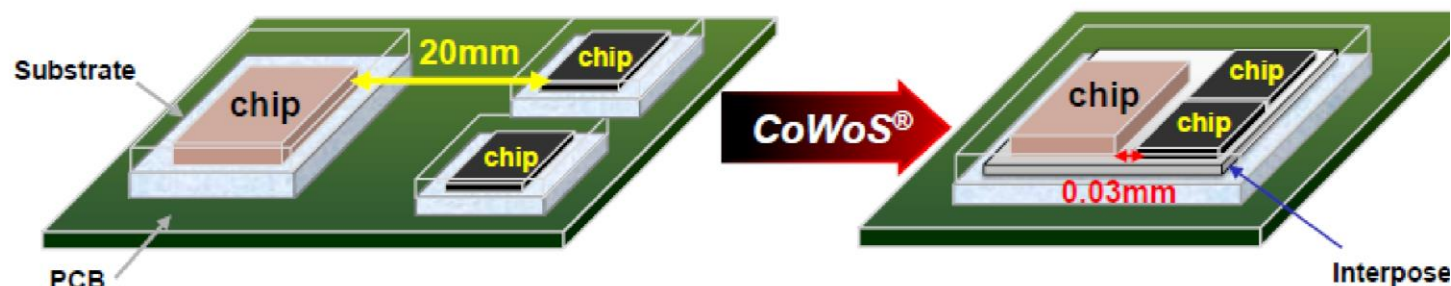
- **Multi-Die InFO :** A solution for homogenous/ heterogeneous integration in a single package



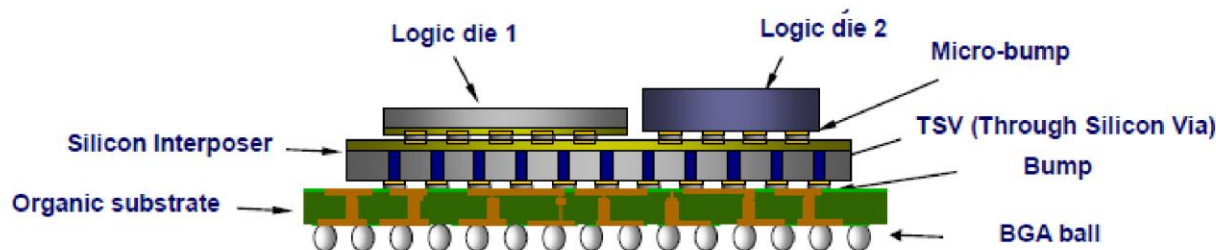
[TSMC 2015]

CoWoS (Chip-On-Wafer-On-Substrate)

- Integrate multiple chips into one single package using a sub-micron scale silicon interface (interposer)



- Enable higher performance, lower power consumption, and smaller form factor
- Best integrated flow for high yield and reliability

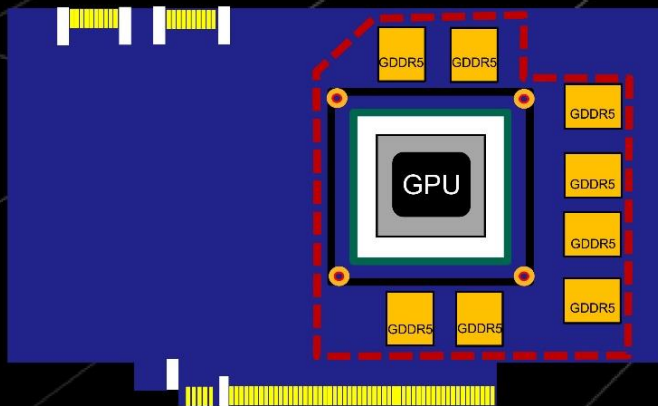


Heterogeneous Integration

[TSMC 2015]

Example of CoWoS

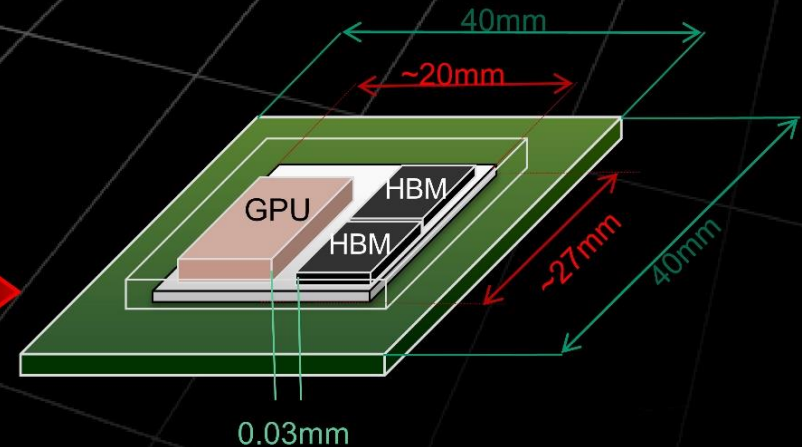
8GByte GDDR5 + GPU on 40x40mm substrate
PCB area ~12cmx12cm
8 GDDR5 BW=192GB/s, Power 42W



High end Graphic Card Example

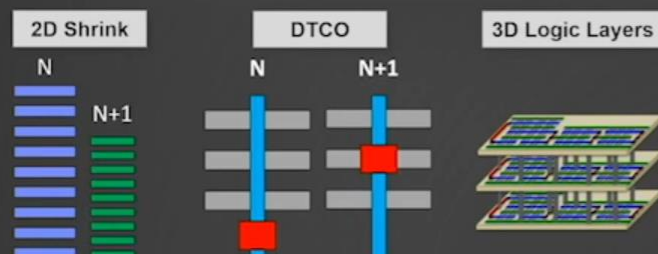
GPU (10x18mm) with 2 HBM stack on interposer 40x40mm substrate
2 HBM, 8 die total, 8GByte, BW 256GB/s, Est. power 7W

CoWoS



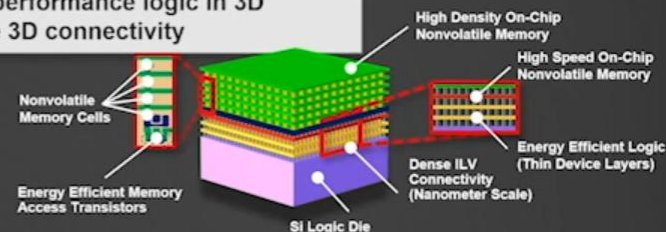
[TSMC 2015]

The Future is **SYSTEM INTEGRATION**



More transistors

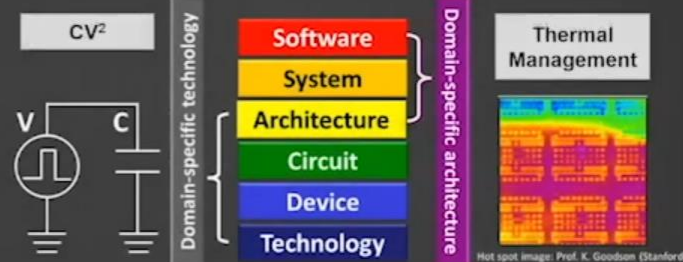
- Local and optimized on-chip memory
- High-performance logic in 3D
- Dense 3D connectivity



More memory



Logic-memory integration



End-to-end optimization

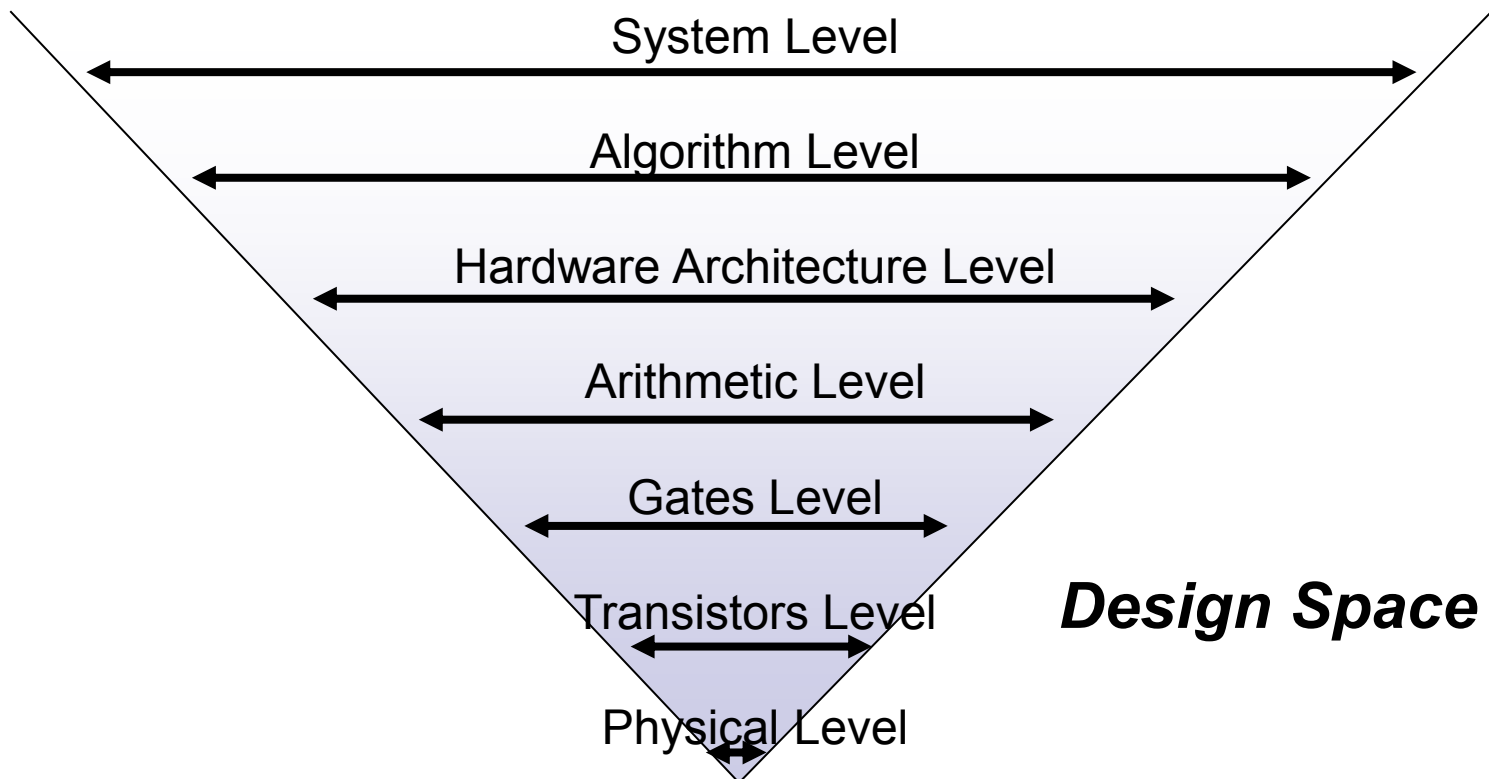
DSP Architecture Design?

- Given DSP algorithms, find the “best” solution in the design space under certain constraints *usually means a good PPA tradeoff.*
- Or, modified or develop the algorithm to be “hardware oriented” or “hardware friendly,” and then develop the hardware architecture
- Domain specific architecture

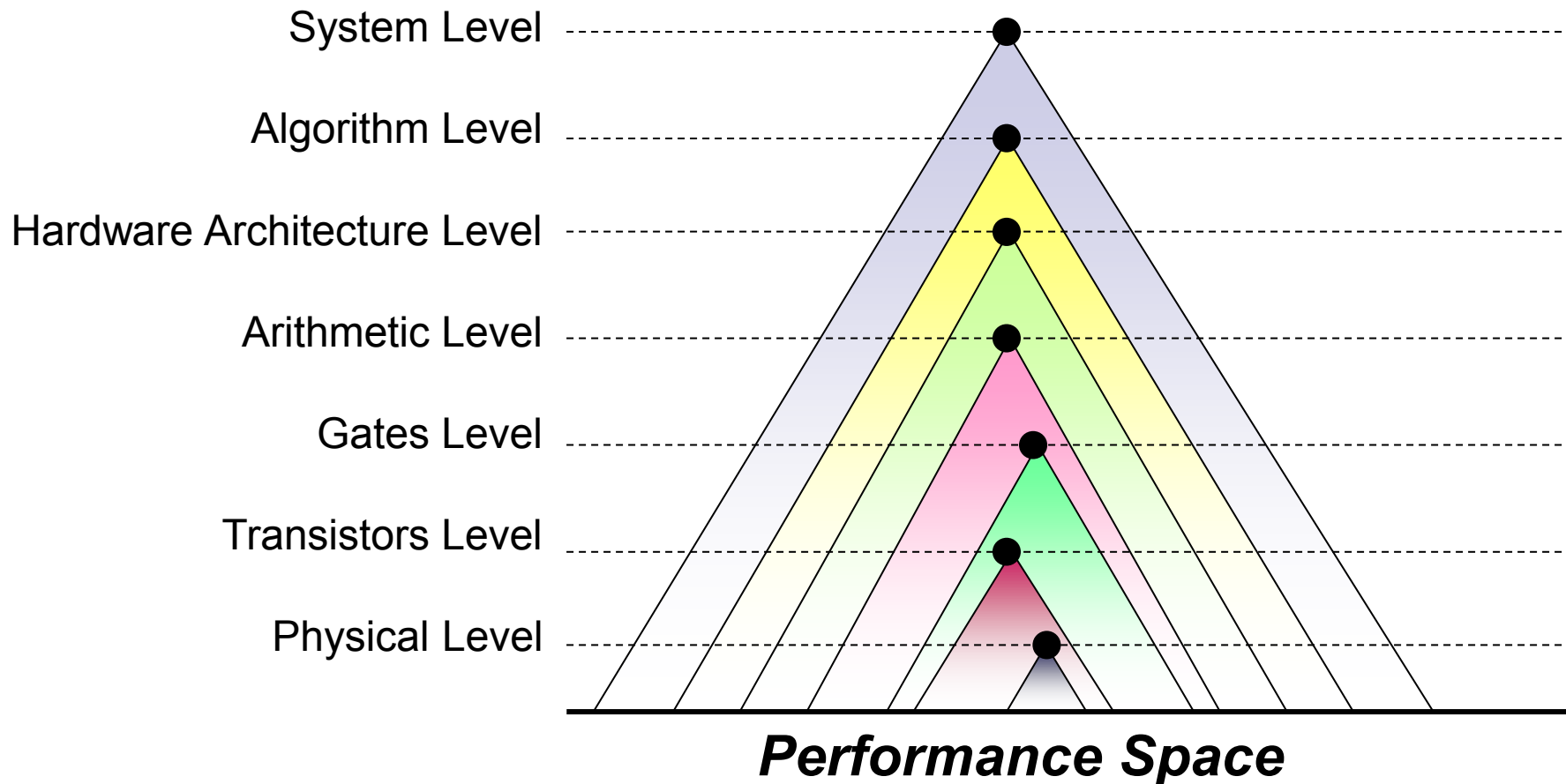
Abstraction Layers

- System (ex: MP3 player)
- Algorithm (ex: FIR filter)
- Hardware architecture (ex: array architecture,...)
- Arithmetic units (ex: multiplier, adder, ...)
- Logic gates (ex: AND, OR, ...)
- Transistors (ex: NMOS, PMOS)
- Layout

The Higher the Abstraction, The Larger Design Space



The Higher the Abstraction, The More Important





Representations of DSP Algorithms

- DSP algorithms: nonterminating program

$$y(n) = ax(n) + bx(n-1) + cx(n-2) \text{ for } n = 1 \text{ to } n = \infty.$$

- Iteration period
- Sampling rate
- Latency
- Throughput
- Clock frequency
- Critical path



Graphical Representations of DSP Algorithms

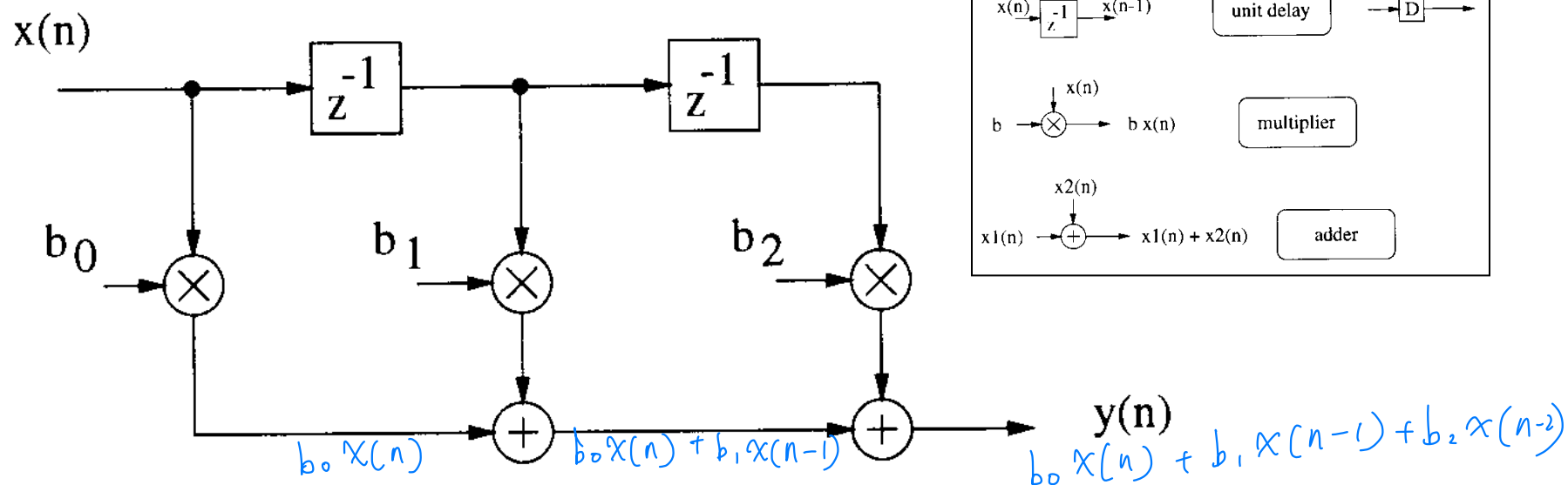
- Can bridge the gap between algorithmic descriptions and structural implementations
- Block diagram
- Signal-flow graph (SFG)
- Data-flow graph (DFG)
- Dependence graph (DG)

Block Diagram (1/5)

- The most frequently used representation
- Can be constructed with different levels of abstraction
- Can be directly mapped to circuits implementation

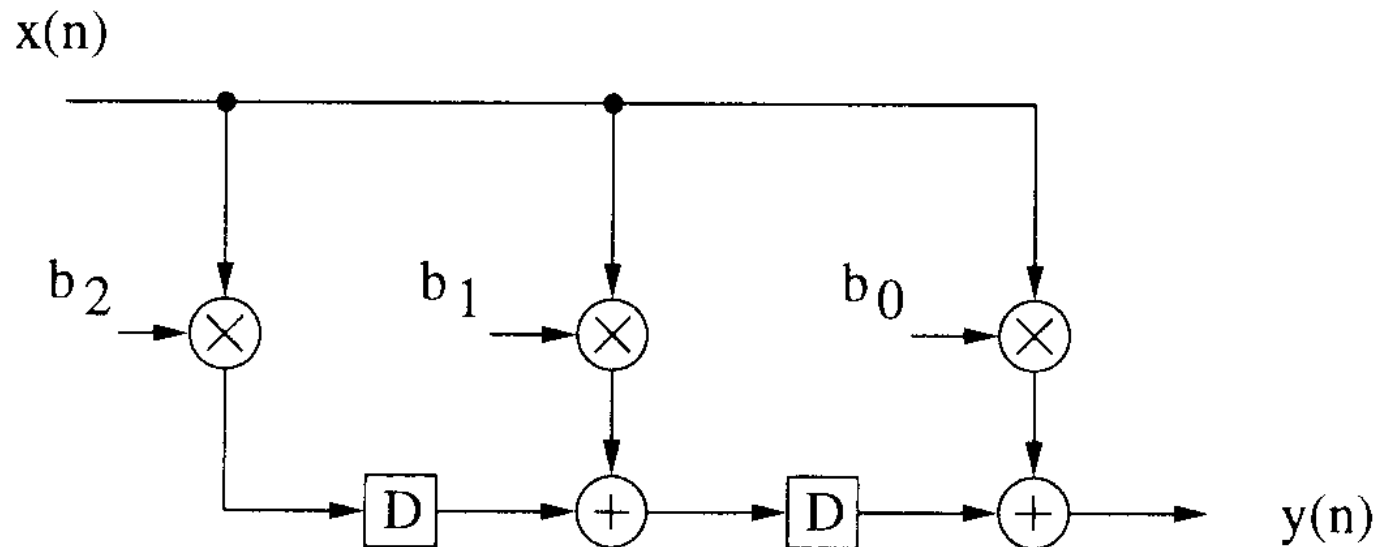
Block Diagram (2/5)

$$y(n] = b_0x(n) + b_1x(n - 1) + b_2x(n - 2)$$

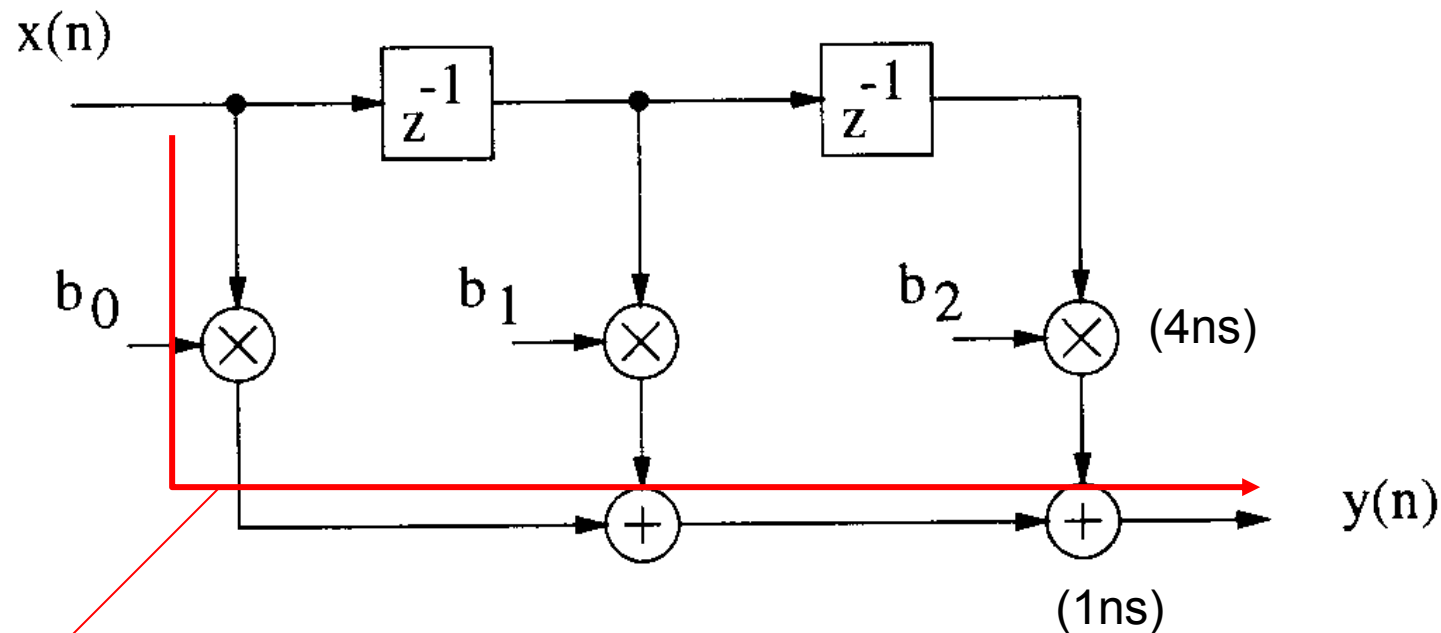


Block Diagram (3/5)

■ Data broadcast FIR filter



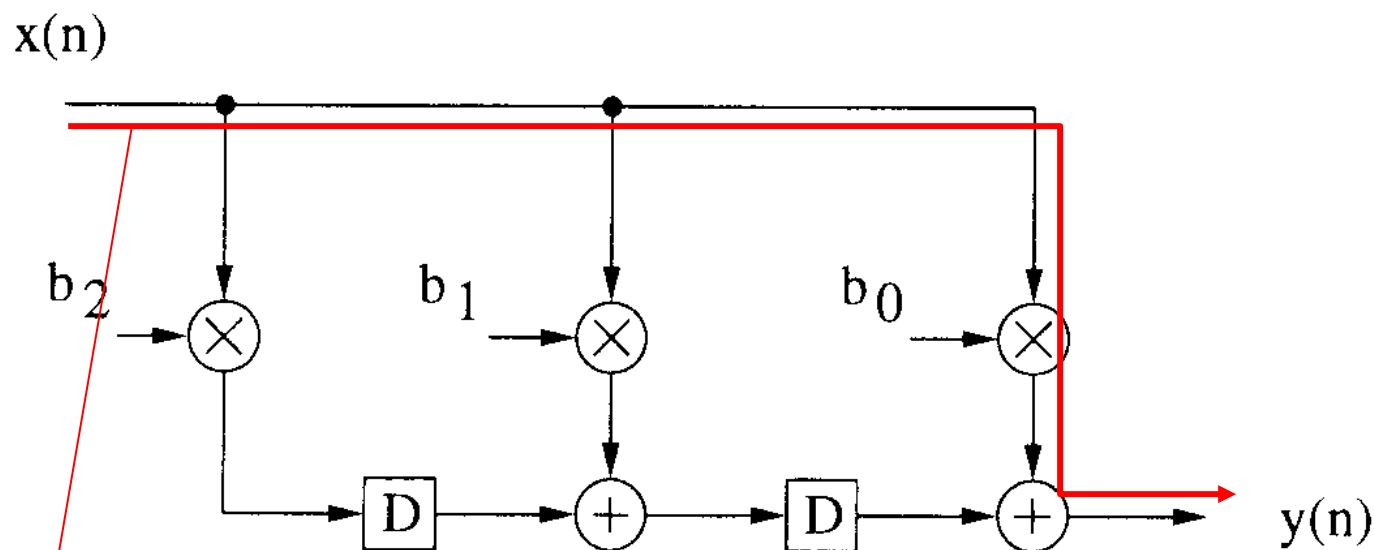
Block Diagram (4/5)



Critical path: $4+1+1=6\text{ns}$

Max clock frequency = $1\text{s}/6\text{ns}=167\text{MHz}$

Block Diagram (5/5)



Critical path: $4+1=5\text{ns}$

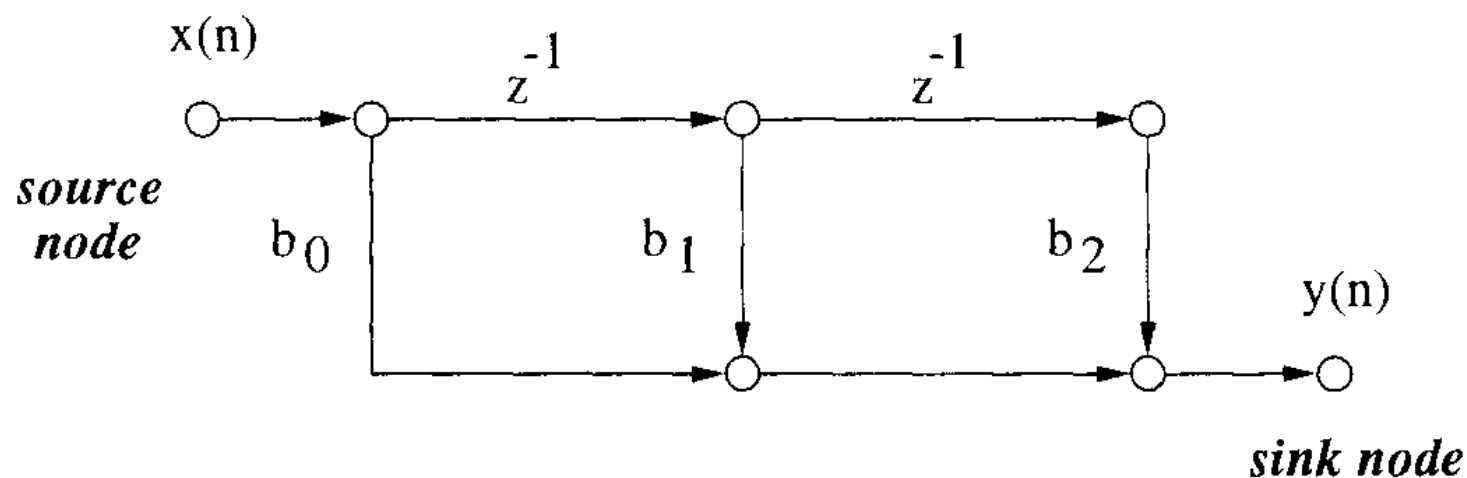
Max clock frequency = $1\text{s}/5\text{ns}=200\text{MHz}$



Signal Flow Graph (SFG) (1/4)

- Nodes k
 - Computation or task
- Directed edges (j, k)
 - Linear transformation
- Source node
- Sink node

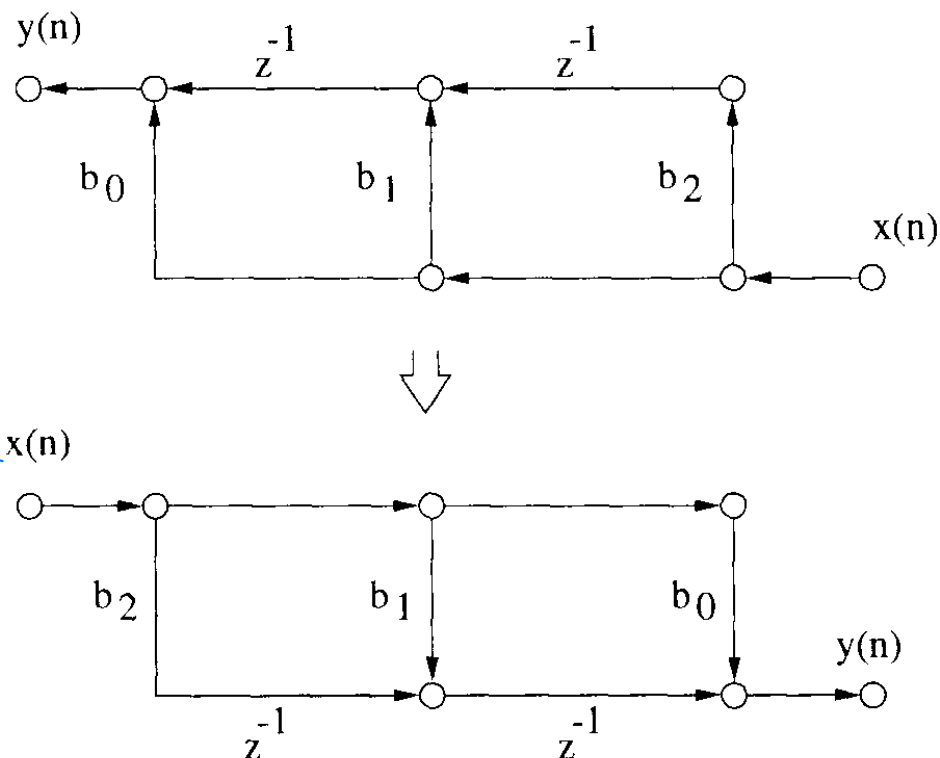
Signal Flow Graph (SFG) (2/4)



Signal Flow Graph (SFG) (3/4)

■ Transpose property

Transposition of an SFG is carried out by reserving the directions of all the edges, exchanging input and output nodes while keeping the edge gain or edge delay unchanged. The resulting SFG maintains the same functional property.





Signal Flow Graph (SFG) (4/4)

- Used in digital filter structure and analysis of finite word-length effects
- Only applicable to linear networks
- Cannot be used to describe multi-rate DSP systems



Data-Flow Graph (DFG) (1/4)

- Nodes

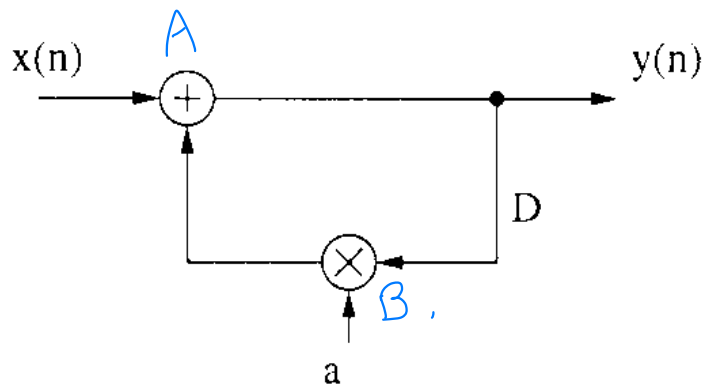
- Computations

- Directed edges

- Data paths (communication)

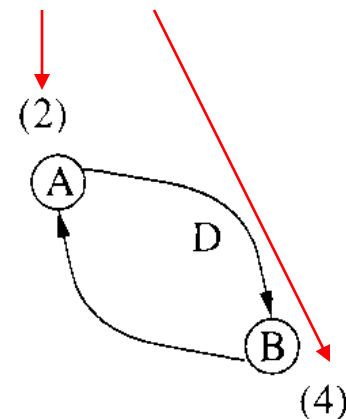
- Has a nonnegative number of delays

Data-Flow Graph (DFG) (2/4)



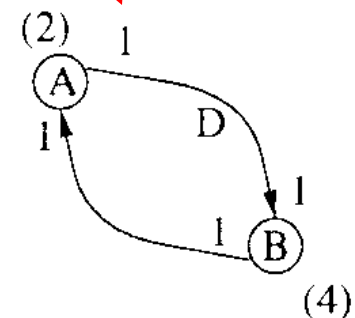
$$y(n) = x(n) + ay(n-1]$$

Execution time



A: +
B: X

Rate



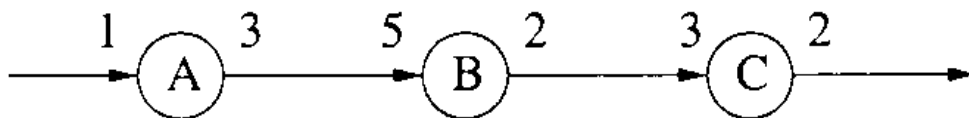
Synchronous DFG

Data-Flow Graph (DFG) (3/4)

- Data-driven property of DSP
 - Any node can fire whenever all the input data are available
 - Intra-iteration precedence constraint
 - Inter-iteration precedence constraint
- Can be used to describe both linear single-rate and nonlinear multi-rate DSP systems

Data-Flow Graph (DFG) (4/4)

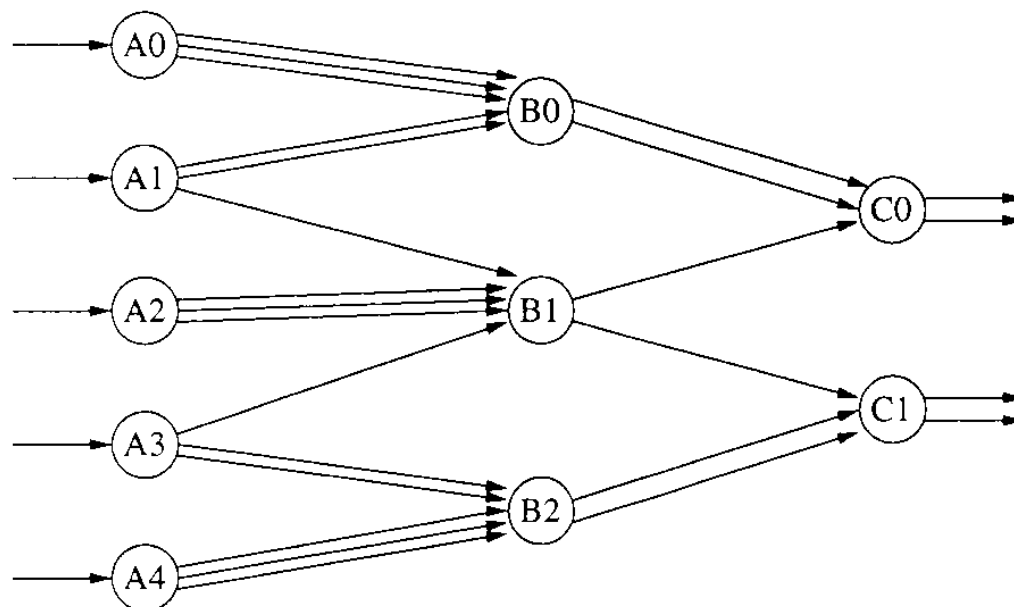
- Use single rate DFG (SRDFG) to represent multi-rate DFG (MRDFG)



$$3f_A = 5f_B$$

$$2f_B = 3f_C$$

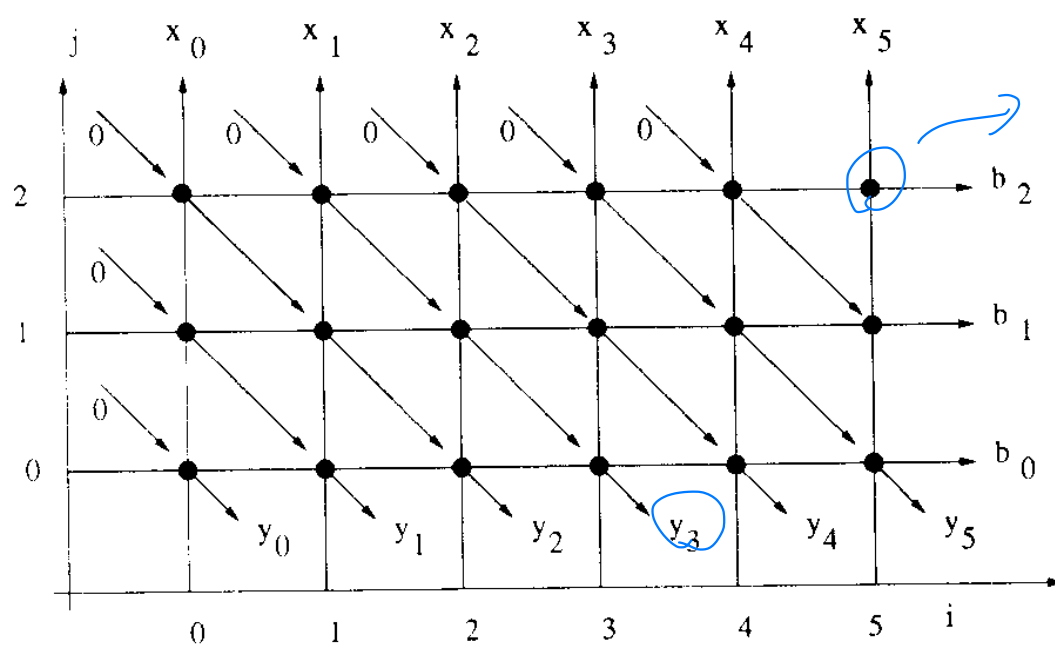
single rate =



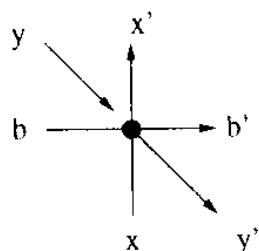
Dependence Graph (1/2)

- A directed graph that shows the dependence of the computation
- Node: computation
- No node in a DG is ever reused on a single computation basis
 - Single-assignment representation (only used - once) -
- Used for systolic-array design

Dependence Graph (2/2)



every could only use once!!



$$\begin{aligned} b' &= b \\ x' &= x \\ y' &= y + b x \end{aligned}$$

$y(3) = b_0 x(3) + b_1 x(2) + b_2 x(1).$

$$y(n) = b_0 x(n) + b_1 x(n-1) + b_2 x(n-2)$$

DFG v.s. DG

■ DFG

- Nodes only cover computation in one iteration, and will be reused iteratively
- Contain delay elements

■ DG

- Contains computation for all iterations, and is used only once
- No delay elements contained

←→
interchangeable