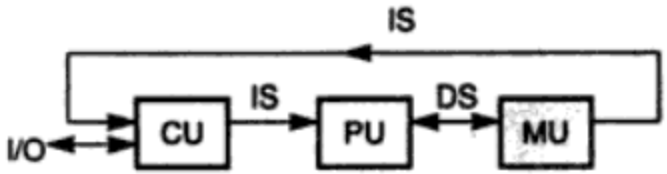
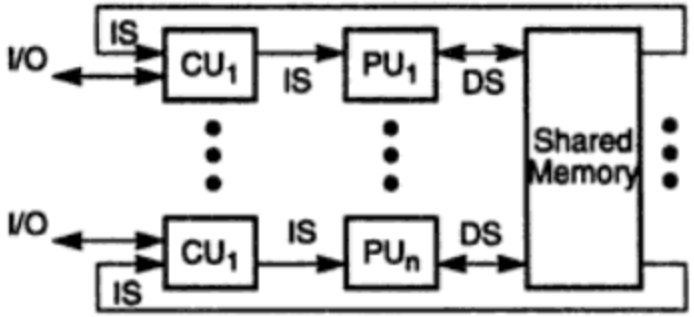
VLSI Models

Ujjwal kumar
Dept. of IT
Gaya College,Gaya

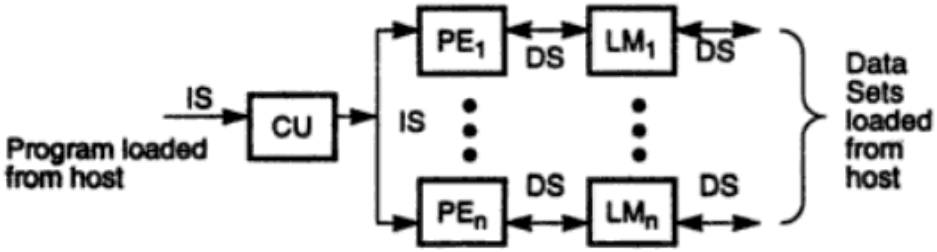
Flynn's Classification



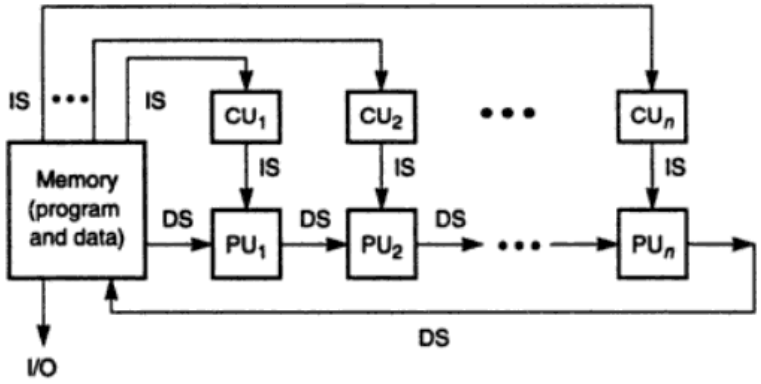
(a) SISD uniprocessor architecture



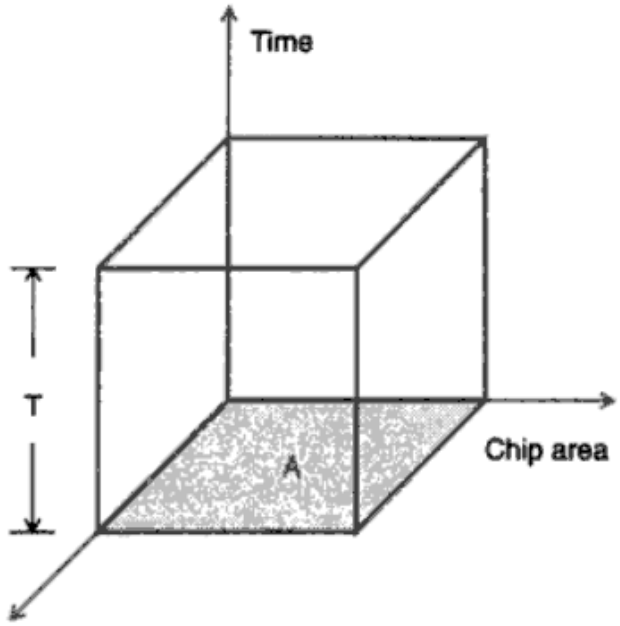
(c) MIMD architecture (with shared memory)



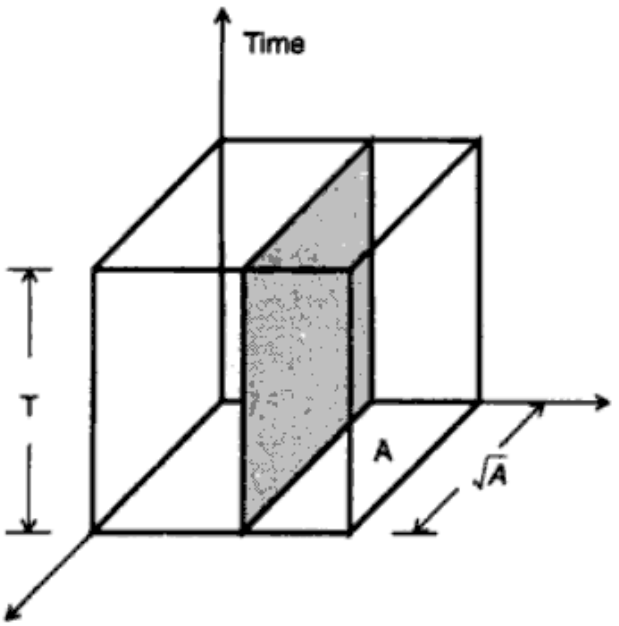
(b) SIMD architecture (with distributed memory)



(d) MISD architecture (the systolic array)



(a) Memory-limited bound on chip area A and I/O-limited bound on chip history represented by the volume AT



(b) Communication-limited bound on the bisection $\sqrt{A}T$

Figure 1.15 The AT^2 complexity model of two-dimensional VLSI chips.

Bisection Communication Bound

- The bisection is represented by the vertical slice cutting across the shorter dimension of the chip area.
- The distance of this dimension is at most AT^2 for a square chip.
- The height of the cross section is T.
- The bisection area represents the maximum amount of information exchange be-tween the two halves of the chip circuit during the time period T.

Bisection Communication Bound

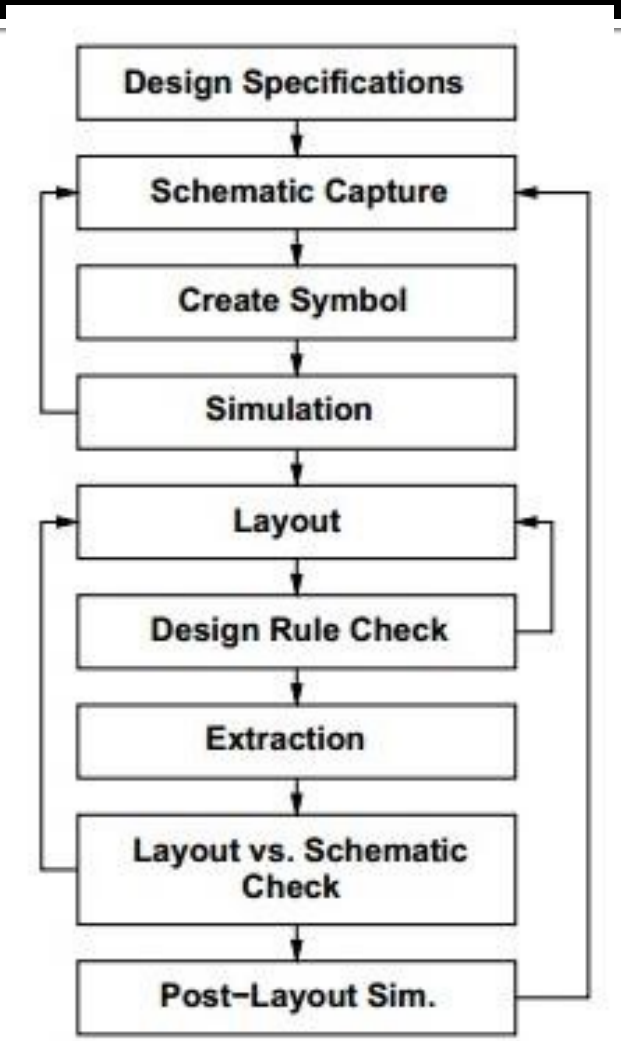
- The cross-section area $(AT)^2$ limits the communication bandwidth of a computation.
- VLSI complexity theoreticians have used the square of this measure, $(AT)^2$, as the lower bound.
- Charles Seitz (1990) has given another interpretation of the $(AT)^2$ result.
- He considers the area-time product AT the cost of a computation, which can be expected to vary as 1/T.
- This implies that the cost of computation for a two-dimensional chip decreases with the execution time allowed.

- When three-dimensional (multilayer) silicon chips are used, Seitz asserted that the cost of computation, as limited by volume-time product, would vary as $1/\sqrt{T}$.
- This is due to the fact that the bisection will vary as $(AT)^{2/3}$ for 3-D chips instead of as $(AT)^2$ for 2-D chips

VLSI Design Flow

- The VLSI IC circuits design flow is shown in the figure below. The various levels of design are numbered and the blocks show processes in the design flow.
- Specifications comes first, they describe abstractly, the functionality, interface, and the architecture of the digital IC circuit to be designed.

- Behavioural description is then created to analyze the design in terms of functionality, performance, compliance to given standards, and other specifications.
- A gate level netlist is a description of the circuit in terms of gates and connections between them, which are made in such a way that they meet the timing, power and area specifications are integrated.
- Finally, a physical layout is made, which will be verified and then sent to fabrication.

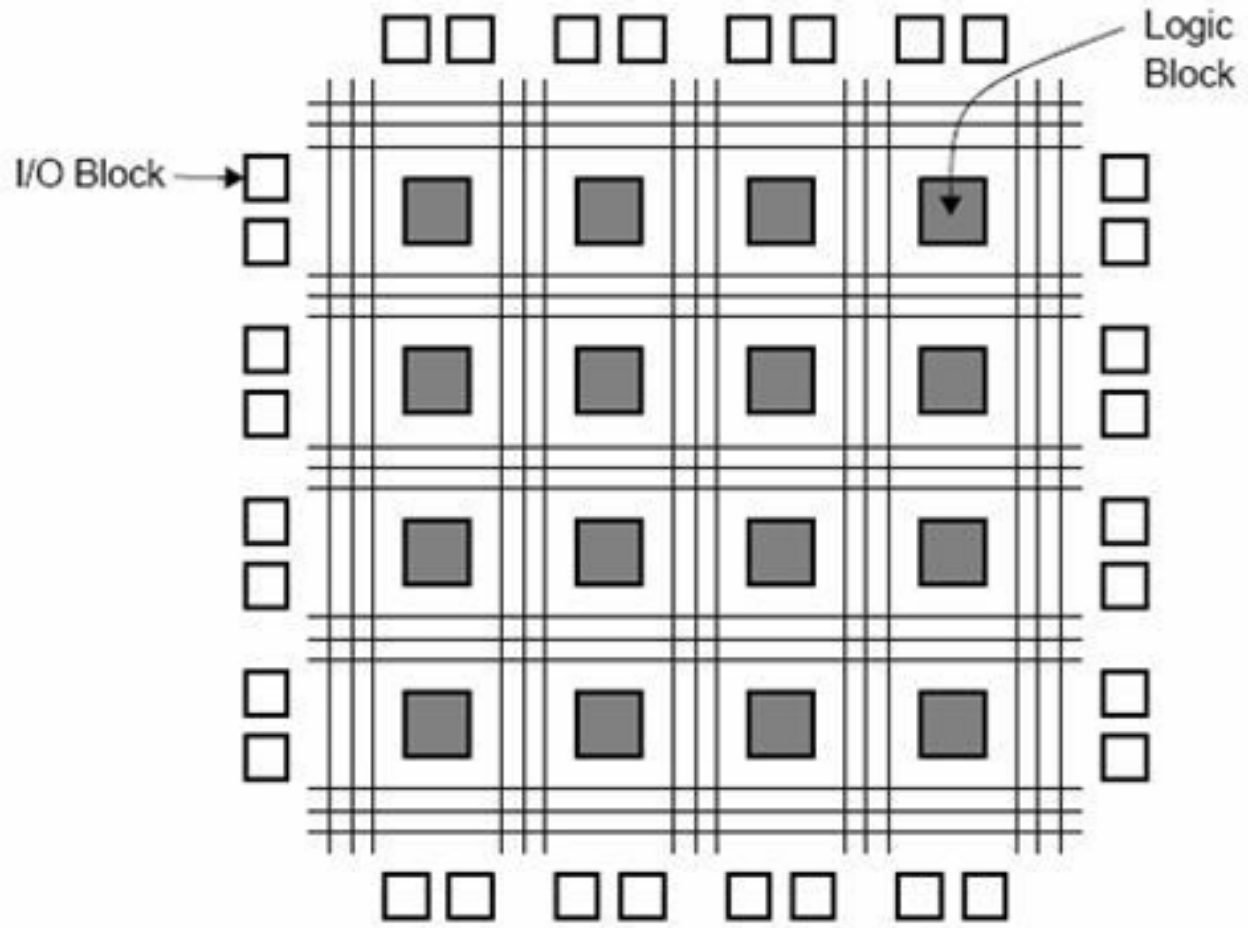

Design Specifications
Schematic Capture
Create Symbol
Simulation
Layout
Design Rule Check
Extraction
Layout vs. Schematic Check
Post-Layout Sim.

VLSI Design - FPGA Technology

- FGPA stands for **Field Programmable Gate Array**
- It contains ten thousand to more than a million logic gates with programmable interconnection.
- Programmable interconnections are available for users or designers to perform given functions easily.
- There are I/O blocks, which are designed and numbered according to function.
- For each module of logic level composition, there are **CLB's (Configurable Logic Blocks)**.
- CLB performs the logic operation given to the module.

VLSI Design - FPGA Technology

- The inter connection between CLB and I/O blocks are made with the help of horizontal routing channels, vertical routing channels and PSM (Programmable Multiplexers).
- The number of CLB it contains only decides the complexity of FPGA. The functionality of CLB's and PSM are designed by VHDL or any other hardware descriptive language. After programming, CLB and PSM are placed on chip and connected with each other with routing channels.


Logic
Block
I/O Block

Advantages

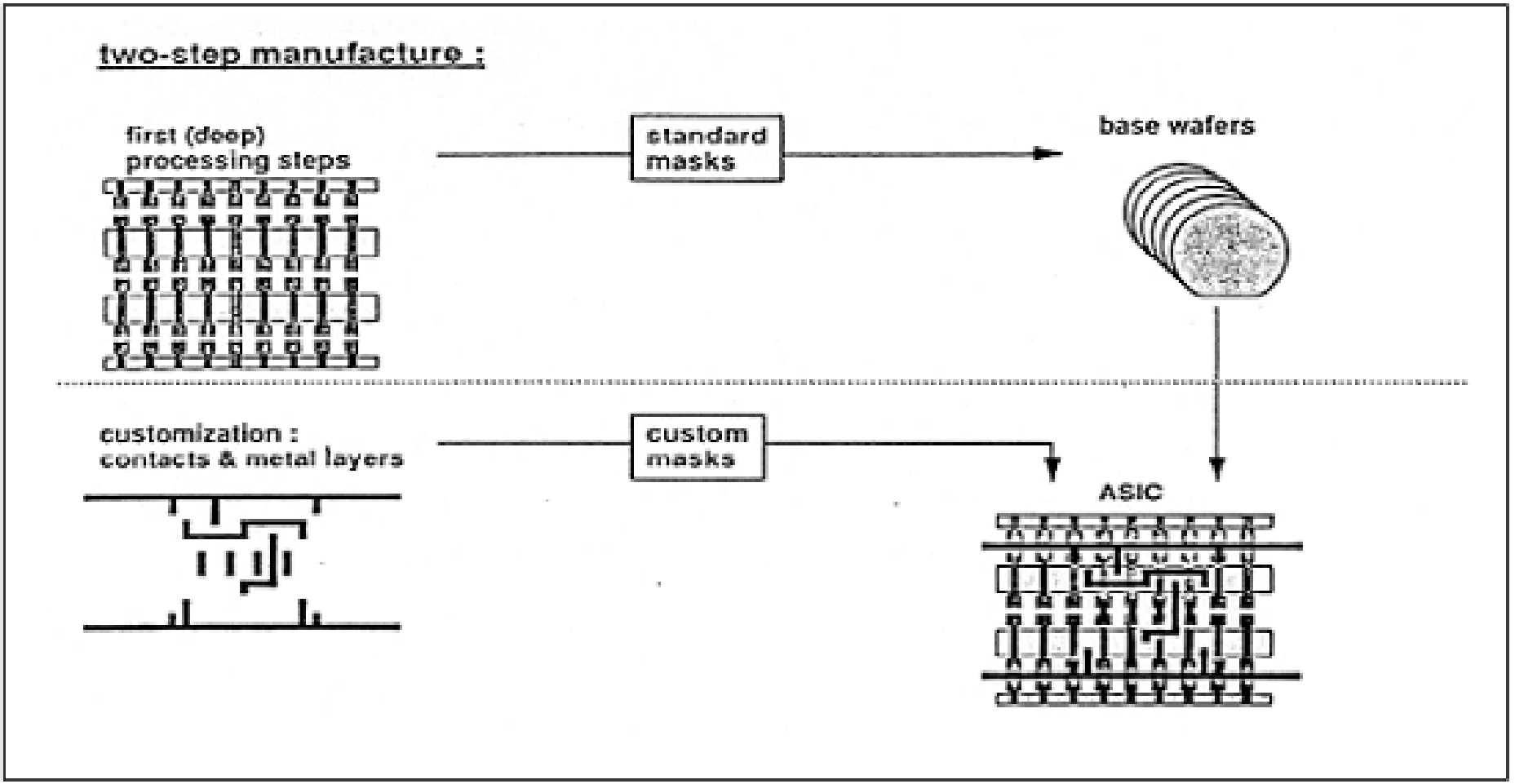
- It requires very small time; starting from design process to functional chip.
- No physical manufacturing steps are involved in it.
- The only disadvantage is, it is costly than other styles.

Gate Array Design

- The **gate array (GA)** ranks second after the FPGA, in terms of fast prototyping capability.
- While user programming is important to the design implementation of the FPGA chip, metal mask design and processing is used for GA.
- Gate array implementation requires a two-step manufacturing process.
  - The first phase results in an array of uncommitted transistors on each GA chip.
  - These uncommitted chips can be stored for later customization, which is completed by defining the metal interconnects between the transistors of the array.
  - The patterning of metallic interconnects is done at the end of the chip fabrication process, so that the turn-around time can still be short, a few days to a few weeks.

- Typical gate array platforms use dedicated areas called channels, for inter-cell routing between rows or columns of MOS transistors.
- They simplify the interconnections. Interconnection patterns that perform basic logic gates are stored in a library, which can then be used to customize rows of uncommitted transistors according to the netlist.

- In most of the modern GAs, multiple metal layers are used for channel routing.
- With the use of multiple interconnected layers, the routing can be achieved over the active cell areas; so that the routing channels can be removed as in Sea-of-Gates (SOG) chips.
- Here, the entire chip surface is covered with uncommitted nMOS and pMOS transistors.
- The neighbouring transistors can be customized using a metal mask to form basic logic gates


two-step manufacture :
first (deep)
processing steps
standard
masks
base wafers
customization :
contacts & metal layers
custom
masks
ASIC

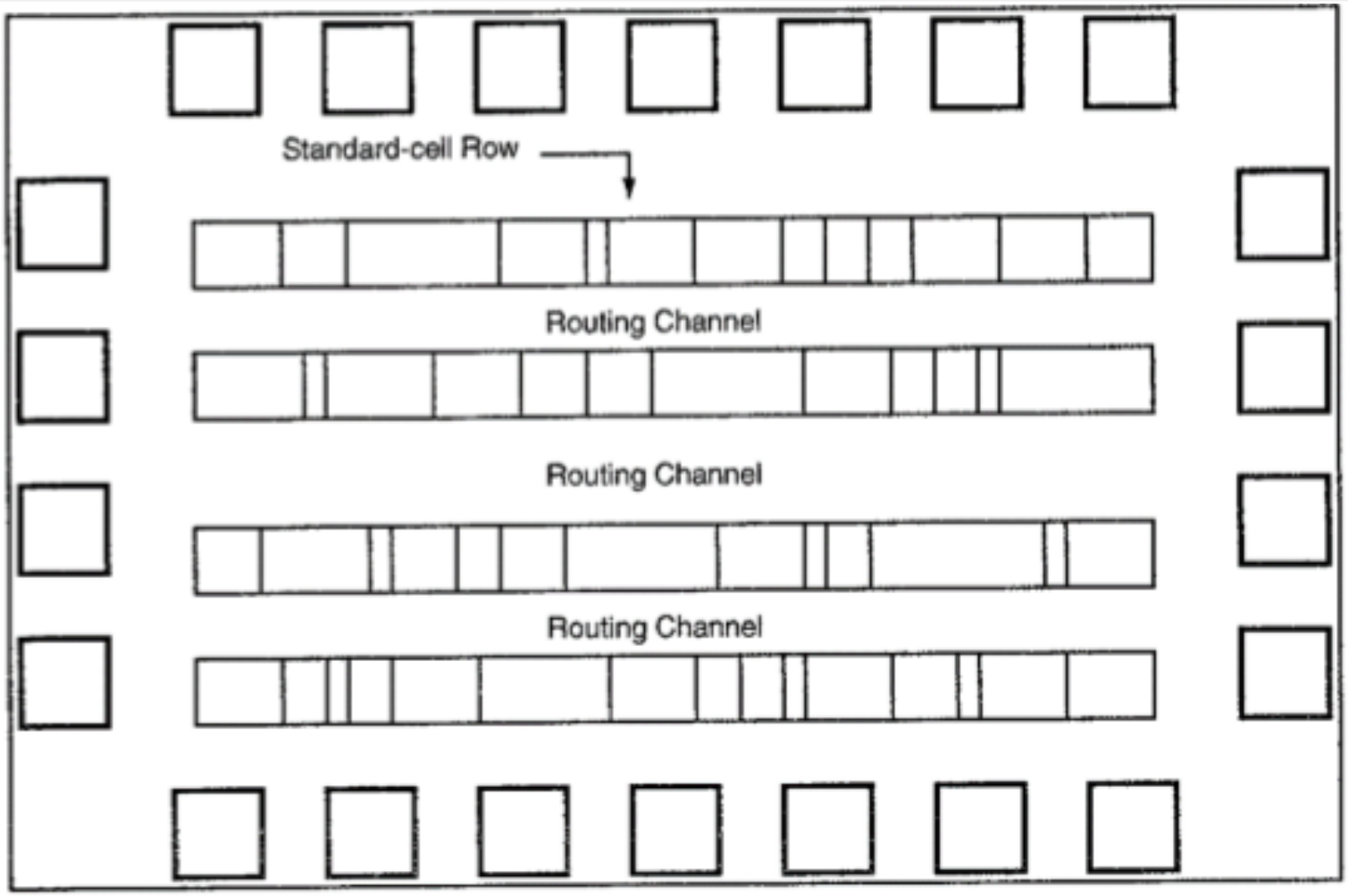
- For inter cell routing, some of the uncommitted transistors must be sacrificed.
- This design style results in more flexibility for interconnections and usually in a higher density.
- GA chip utilization factor is measured by the used chip area divided by the total chip area.
- It is higher than that of the FPGA and so is the chip speed.

Standard Cell Based Design

- A standard cell based design requires development of a full custom mask set.
- The standard cell is also known as the polycell.
- In this approach, all of the commonly used logic cells are developed, characterized and stored in a standard cell library.
- A library may contain a few hundred cells including inverters, NAND gates, NOR gates, complex AOI, OAI gates, D-latches and Flip-flops.
- Each gate type can be implemented in several versions to provide adequate driving capability for different fan-outs.
- The inverter gate can have standard size, double size, and quadruple size so that the chip designer can select the proper size to obtain high circuit speed and layout density.

- Each cell is characterized according to several different characterization categories, such as,
  - Delay time versus load capacitance
  - Circuit simulation model
  - Timing simulation model
  - Fault simulation model
  - Cell data for place-and-route
  - Mask data

- For automated placement of the cells and routing, each cell layout is designed with a fixed height, so that a number of cells can be bounded side-by-side to form rows.
- The power and ground rails run parallel to the upper and lower boundaries of the cell.
- So that, neighbouring cells share a common power bus and a common ground bus.


Standard-cell Row
Routing Channel
Routing Channel
Routing Channel

Full Custom Design

- In a full-custom design, the entire mask design is made new, without the use of any library.
- The development cost of this design style is rising.
- Thus, the concept of design reuse is becoming famous to reduce design cycle time and development cost.
- The hardest full custom design can be the design of a memory cell, be it static or dynamic.
- For logic chip design, a good negotiation can be obtained using a combination of different design styles on the same chip, i.e. standard cells, data-path cells, and **programmable logic arrays (PLAs)**.

- Practically, the designer does the full custom layout, i.e. the geometry, orientation, and placement of every transistor.
- The design productivity is usually very low; typically a few tens of transistors per day, per designer.
- In digital CMOS VLSI, full-custom design is hardly used due to the high labour cost.
- These design styles include the design of high-volume products such as memory chips, high-performance microprocessors and FPGA.

Architectural Development Tracks

- The evolution of parallel computers I spread along the following tracks –
- Multiple Processor Tracks
  - Multiprocessor track (Shared Memory Track)
  - Multicomputer track (Message Passing Track)
- Multiple data track
  - Vector track
  - SIMD track
- Multiple threads track
  - Multithreaded track
  - Dataflow track

Multiple Processor Tracks

- In **multiple processor track**, it is assumed that different threads execute concurrently on different processors and communicate through shared memory (multiprocessor track) or message passing (multicomputer track) system.
  - Multiprocessor track (Shared Memory Track)
  - Multicomputer track (Message Passing Track)

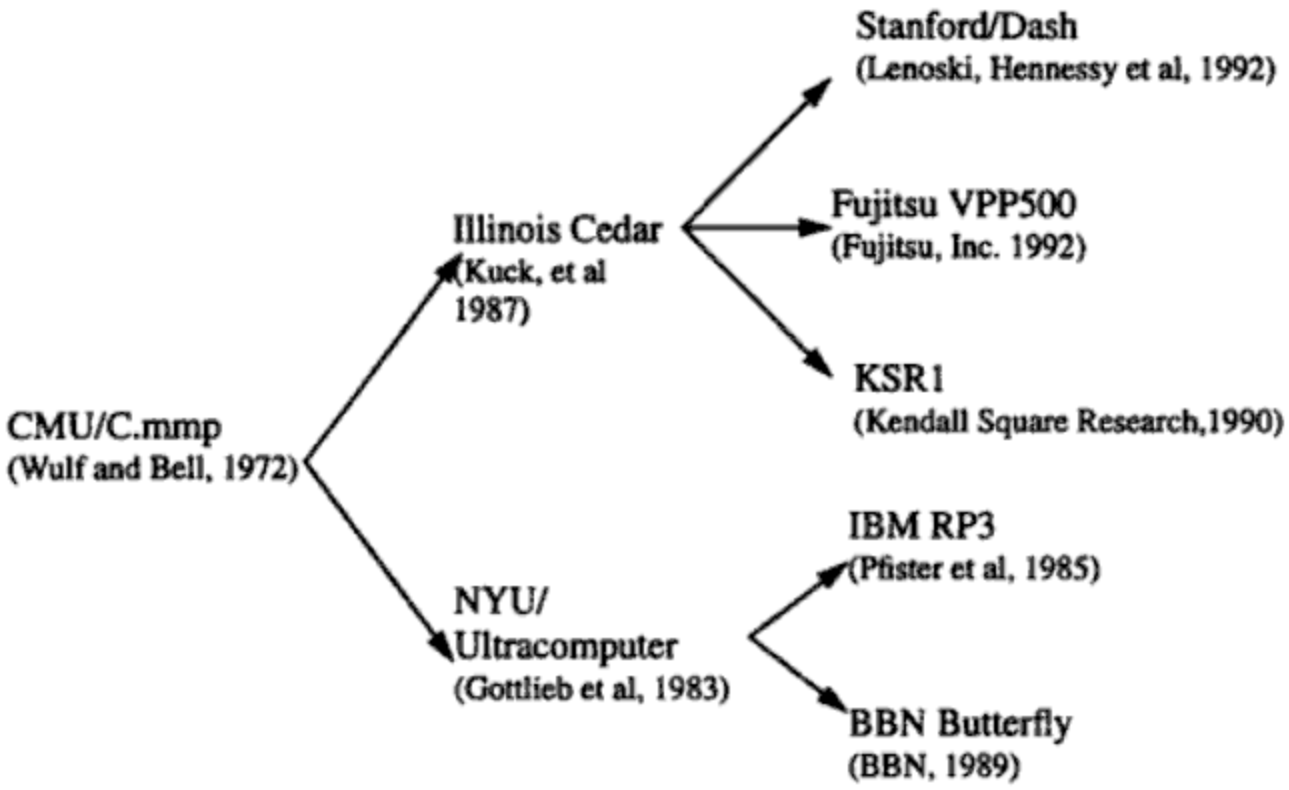
Shared-Memory Track

- The track started with the C.mmp system developed at Carnegie-Mellon University (Wulf and Bell, 1972).
- The C.mmp was an UMA multiprocessor. Sixteen PDP 11/40 processors are interconnected to 16 shared-memory modules via a crossbar switch.
- A special interprocessor interrupt bus is provided for fast interprocess communication, besides the shared memory.
- The C.mmp project pioneered shared-memory multiprocessor development, not only in the crossbar architecture but also in the multiprocessor operating system (Hydra) development.
- Both the NYU Ultracomputer project (Gottlieb et al., 1983) and the Illinois Cedar project (Kuck et al., 1987) were developed with a single address space.

Shared-Memory Track

- Both systems used multistage networks as a system interconnect.
- The major achievements in the Cedar project are in parallel compilers and performance benchmarking experiments.
- The Stanford Dash (Lenoski, Hennessy et al., 1992) is a NUMA multiprocessor with distributed memories forming a global address space.
- Cache coherence is enforced with distributed directories.
- The KSR-1 is a typical COMA model.
- The Fujitsu VPP 500 is a 222-processor system with a crossbar interconnect.
- The shared memories are distributed to all processor nodes.

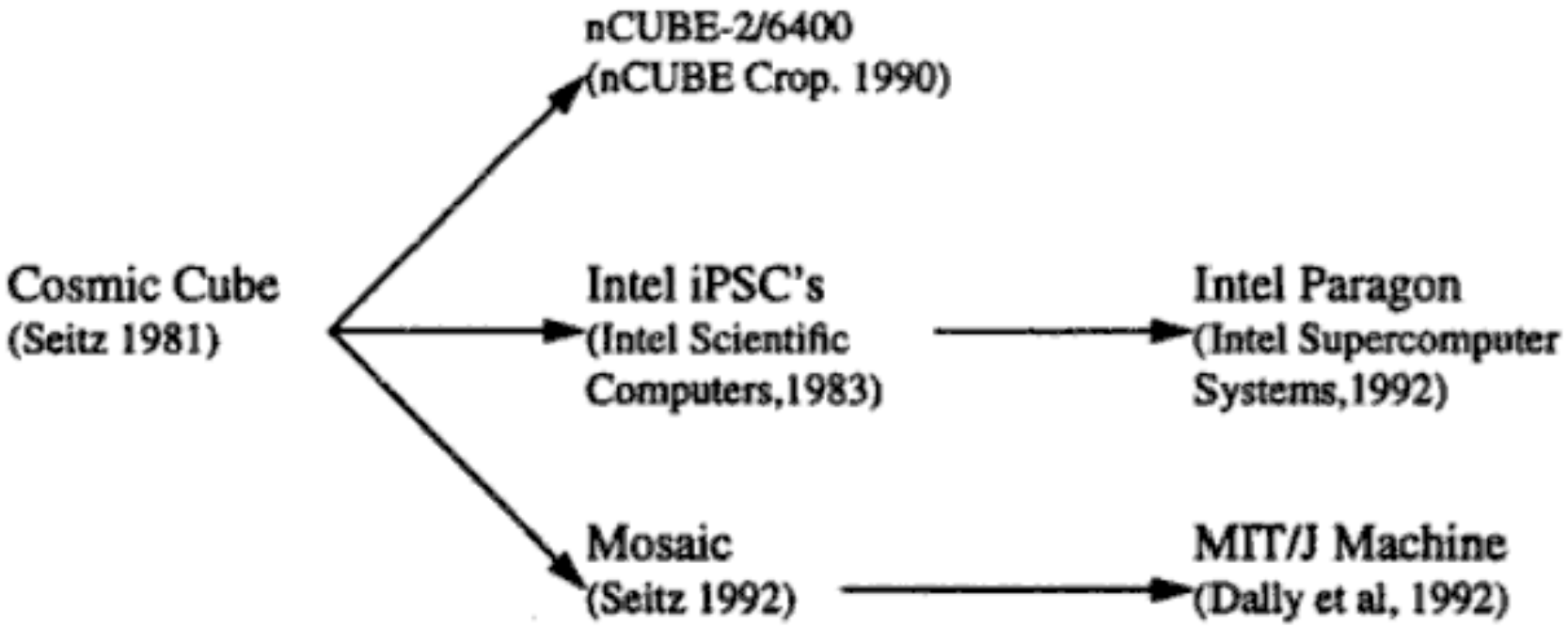
Shared-Memory Track



Message Passing Track

- The Cosmic Cube (Seitz et al., 1981) pioneered the development of message-passing multicomputers.
- Since then, Intel has produced a series of medium-grain hypercube computers (the iPSCs).
- The nCUBE 2 also assumes a hypercube configuration.
- On the research track, the Mosaic C (Seitz, 1992) and the MIT 3-Machine (Daily et al., 1992) are two fine-grain multicomputers.

Message Passing Track



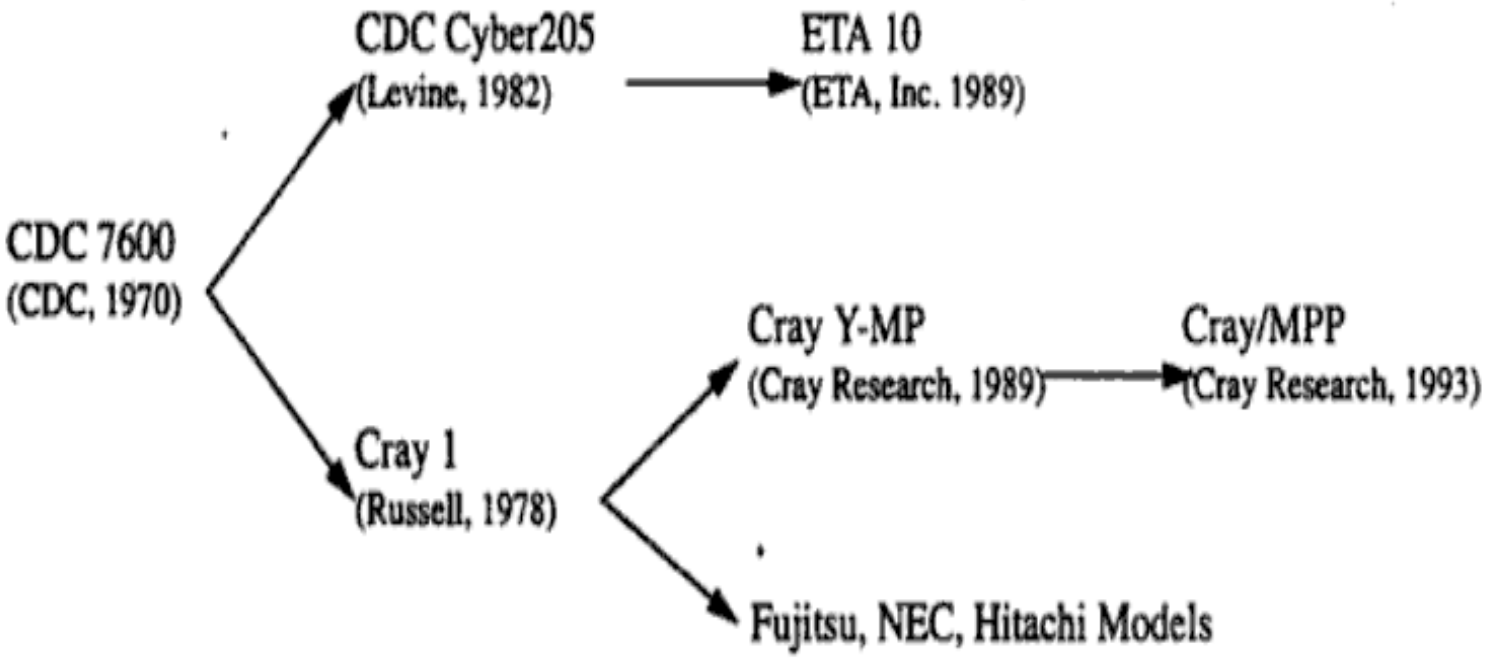
Multiple data track

- In **multiple data track**, it is assumed that the same code is executed on the massive amount of data. It is done by executing same instructions on a sequence of data elements (vector track) or through the execution of same sequence of instructions on a similar set of data (SIMD track).
  - Vector track
  - SIMD track

Multivector Track

- These are traditional vector supercomputers.
- The CDC 7600 was the first vector dual-processor system.
- Two subtracks were derived from the CDC 7600.
- The Cray and Japanese supercomputers all followed the register-to-register architecture.
- Cray 1 pioneered the multivector development in 1978.
- The latest Cray/MPP is a massively parallel system with distributed shared memory.
- It is supposed to work as a back-end accelerator engine compatible with the existing Cray Y-MP Series.
- The other subtrack used memory-to-memory architecture in building vector super-computers.

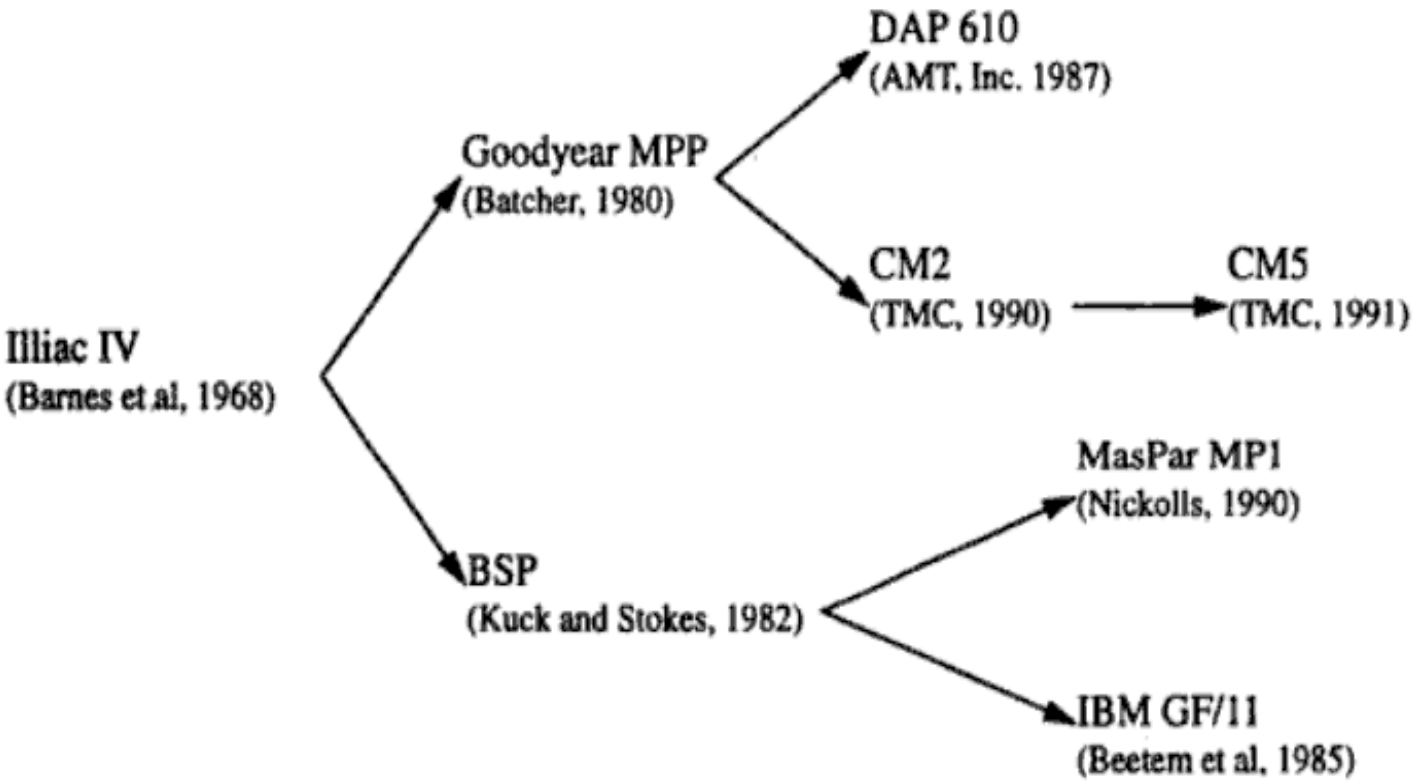
Multivector Track



The SIMD Track

- The ILLIAC IV pioneered the construction of SIMD computers, even the array processor concept can be traced back far earlier to the 1960s.
- The subtrack, consisting of the Goodyear MPP, the AMT/DAP610, and the TMC/CM-2, are all SIMD machines built with bit-slice PEs.
- The CM-5 is a synchronized MIMD machine executing in a multiple-SIMD mode.
- The other subtrack corresponds to medium-grain SIMD computers using word-wide PEs.
- The BSP (Kuck and Stokes, 1982) was a shared-memory SIMD machine built with 16 processors updating a group of 17 memory modules synchronously.
- The GF11 (Beetem et al.. 1985) was developed at the IBM Watson Laboratory for scientific simulation research use.
- The MasPar MP1 is the only medium-grain SIMD computer currently in production use.

The SIMD Track



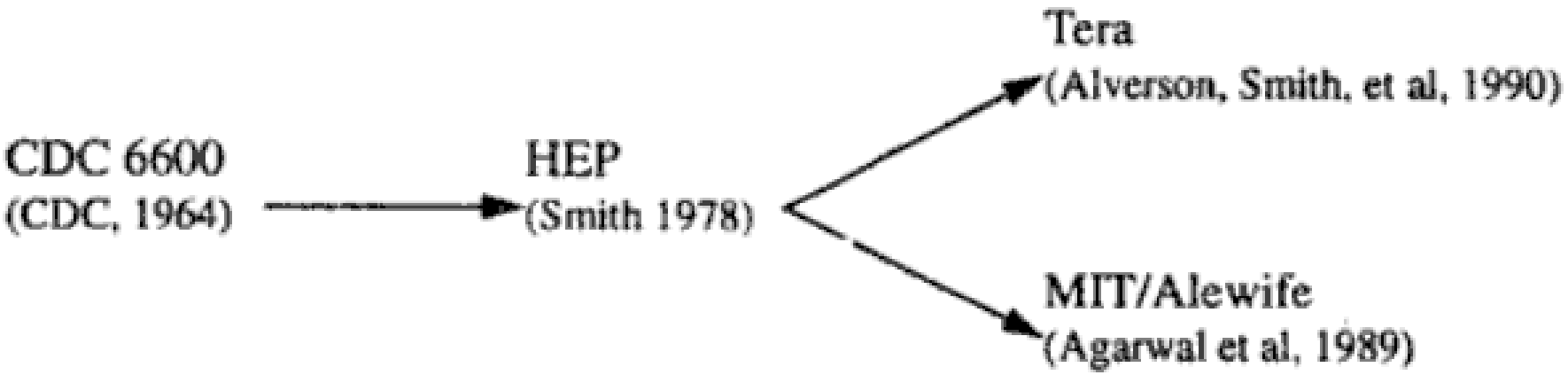
Multiple threads track

- In **multiple threads track**, it is assumed that the interleaved execution of various threads on the same processor to hide synchronization delays among threads executing on different processors. Thread interleaving can be coarse (multithreaded track) or fine (dataflow track).
  - Multithreaded track
  - Dataflow track

Multithreaded Tracks

- The conventional von Neumann machines are built with processors that execute a single context by each processor at a time.
- In other words, each processor maintains a single thread of control with limited hardware resources.
- In a multithreaded architecture. each processor can execute multiple contexts at the same time.
- The term multithreading implies that there are multiple threads of control in each processor.
- Multithreading offers an effective mechanism for hiding long latency in building large-scale multiprocessors.
- The multithreading idea was pioneered by Burton Smith (1978) in the HEP system which extended the concept of score boarding of multiple functional units in the CDC 6400.

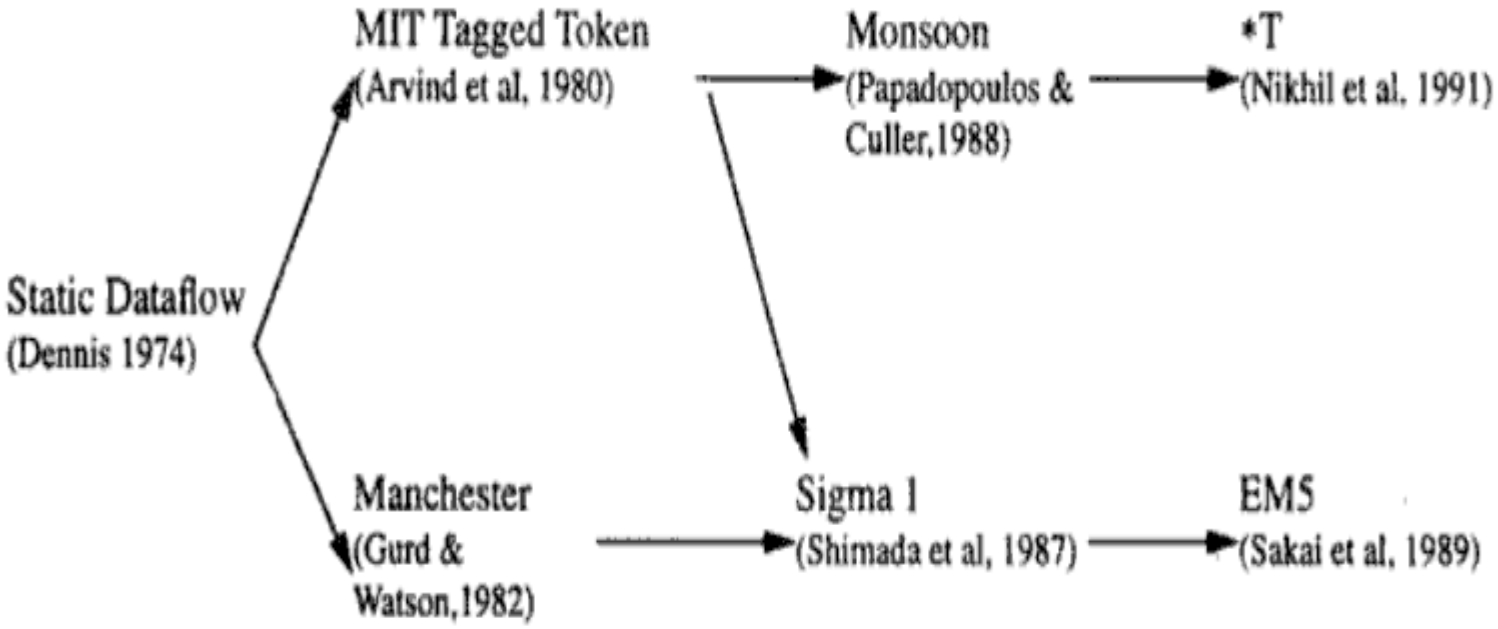
Multithreaded Tracks



The Dataflow Track

- The key idea is to use a dataflow mechanism, instead of a control-flow mechanism as in von Neumann machines, to direct the program flow.
- Fine-grain, instruction-level parallelism is exploited in dataflow computers.
- As listed in Figure, the dataflow concept was pioneered by Jack Dennis (1974) with a "static" architecture.
- The concept later inspired the development of "dynamic" dataflow computers.
- A series of tagged-token architectures was developed at MIT by Arvind and coworkers.
- Then the *T prototype (Nikhil et al., 1991)
- Another important subtrack of dynamic dataflow computer is represented by the Manchester machine (Curd and Watson, 1982).
- The ETL Sigma 1 (Shimada et al, 1987) and EM5 have evolved from the MIT and Manchester machines.

The Dataflow Track



Thank You !