Beyond Exascale:

Extreme-Scale Architecture in the Neo-Digital Age

Thomas Sterling

Professor of Informatics and Computing

Chief Scientist, CREST

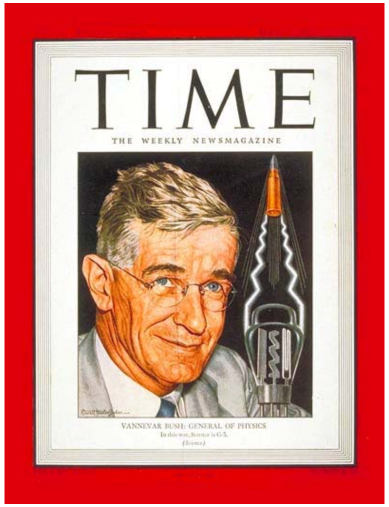
Indiana University

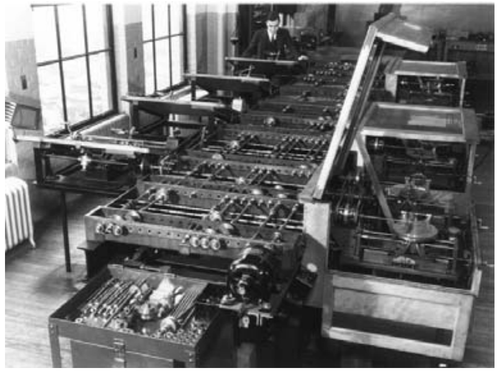
July 9, 2014


NCC-1701-D
THE LIGHT WORKS

Shifting Paradigms of Computing

- Abacus
  - Counting tables
- Pascaline
- Difference engine
  - Charles Babbage
  - Per Georg Scheutz
- Tabulators
  - Herman Hollerith
- Analog computer
  - Vannevar Bush Machine
- Harvard Architecture
  - Howard Aiken
  - Konrad Zuse
  - Charles Babbage Analytical Engine

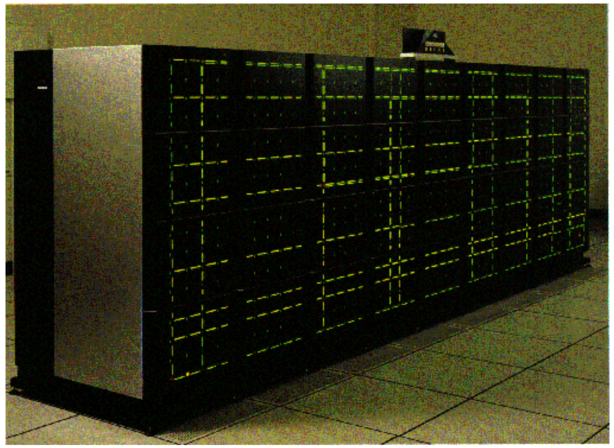




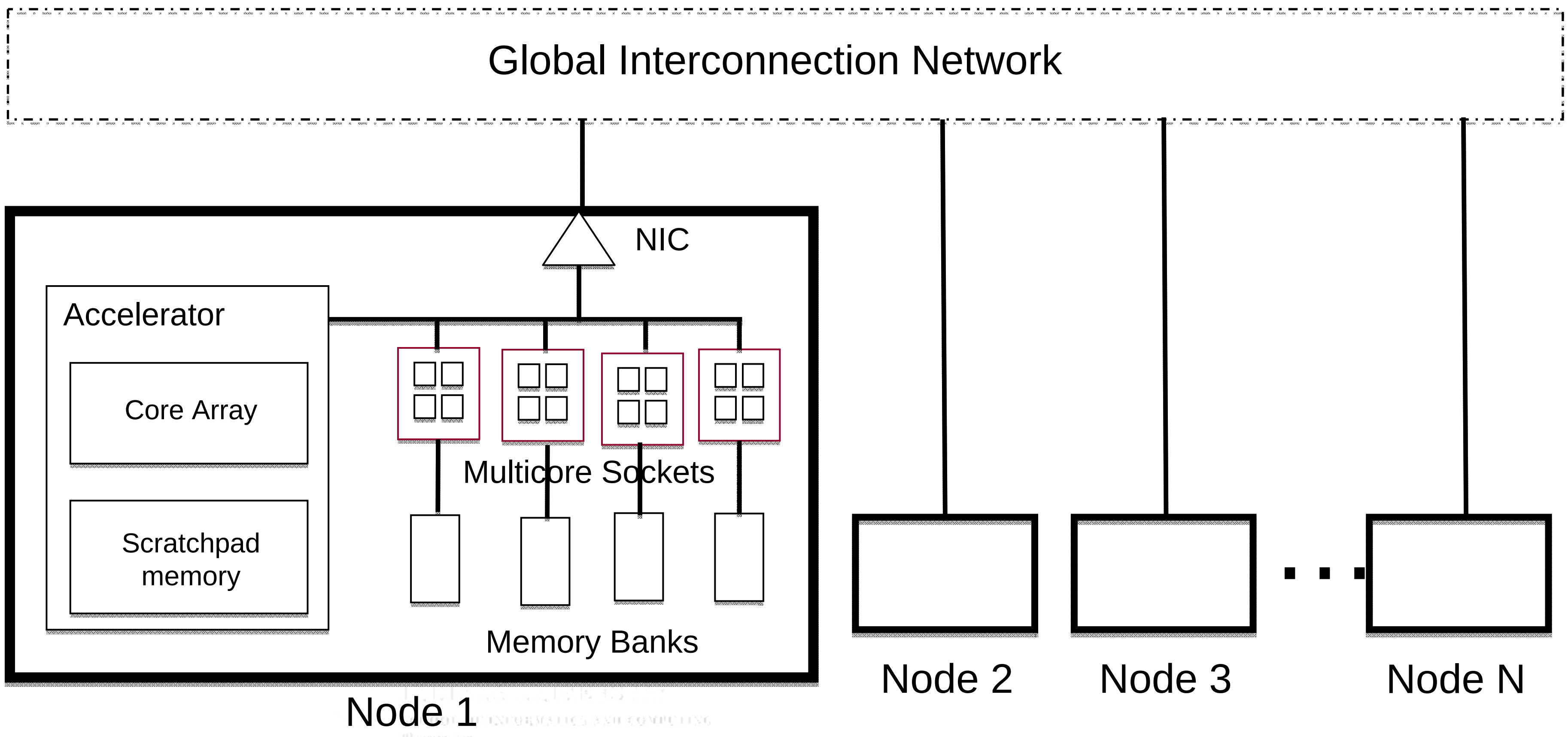
The von Neumann Age

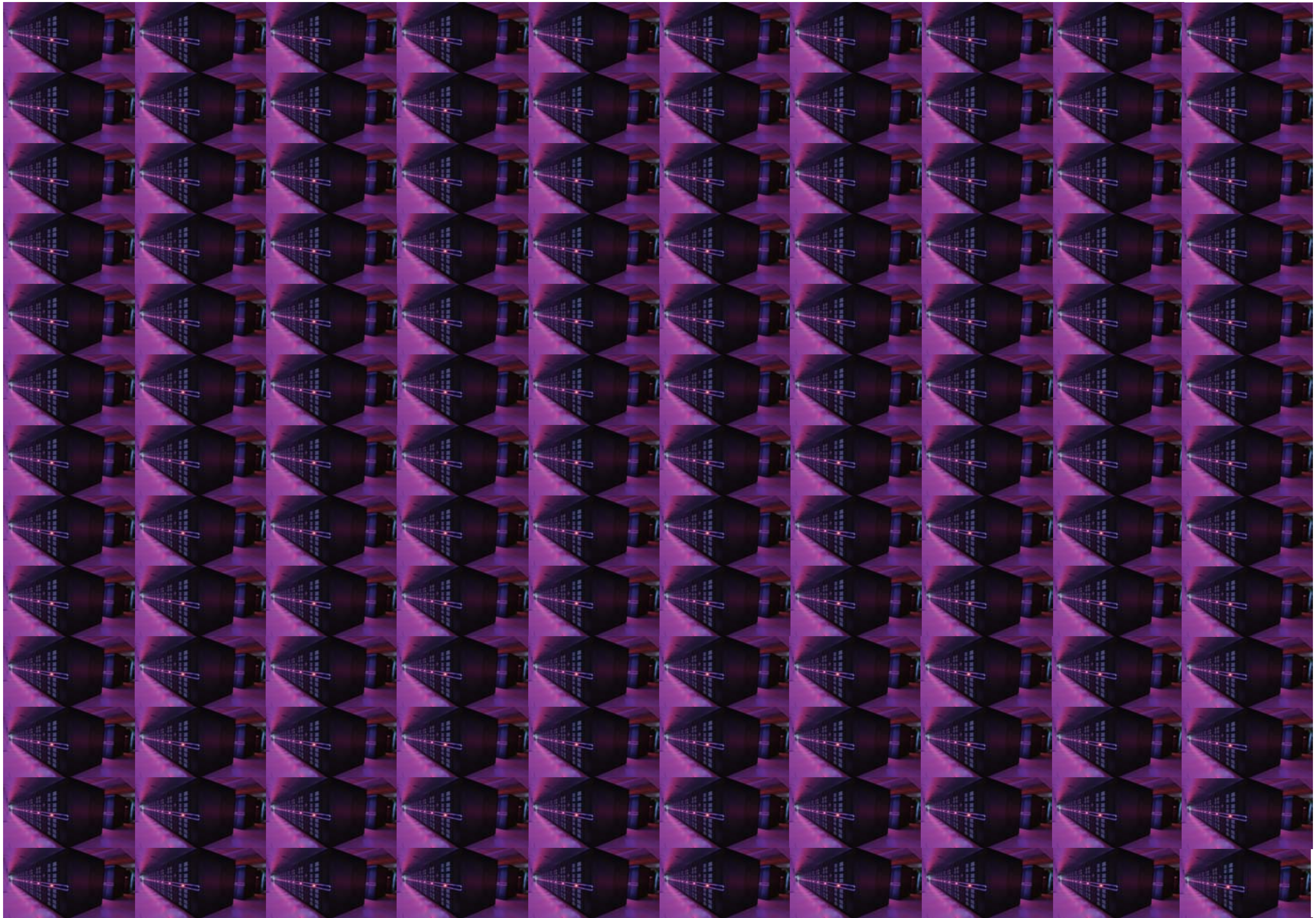
- Foundations:
  – Information Theory – Claude Shannon
  – Computabilty – Turing/Church
  – Cybernetics – Norbert Wiener
  – Stored Program Computer Architecture – von Neumann

- The von Neumann Shift: 1945 – 1960
  – Vacuum tubes, core memory
  – Technology assumptions
    - ALUs are most expensive components
    - Memory capacity and clock rate are scale drivers – mainframes
    - Data movement of secondary importance
- Von Neumann extended: 1960 – 2014
  – Semiconductor, Exploitation of parallelism
  – Out of order completion
  – Vector
  – SIMD
  – Multiprocessor (MIMD)
    - SMP
      – Maintain sequential consistency
    - MPP/Clusters
      – Ensemble computations with message passing



Conventional Heterogeneous Multicore System Architecture

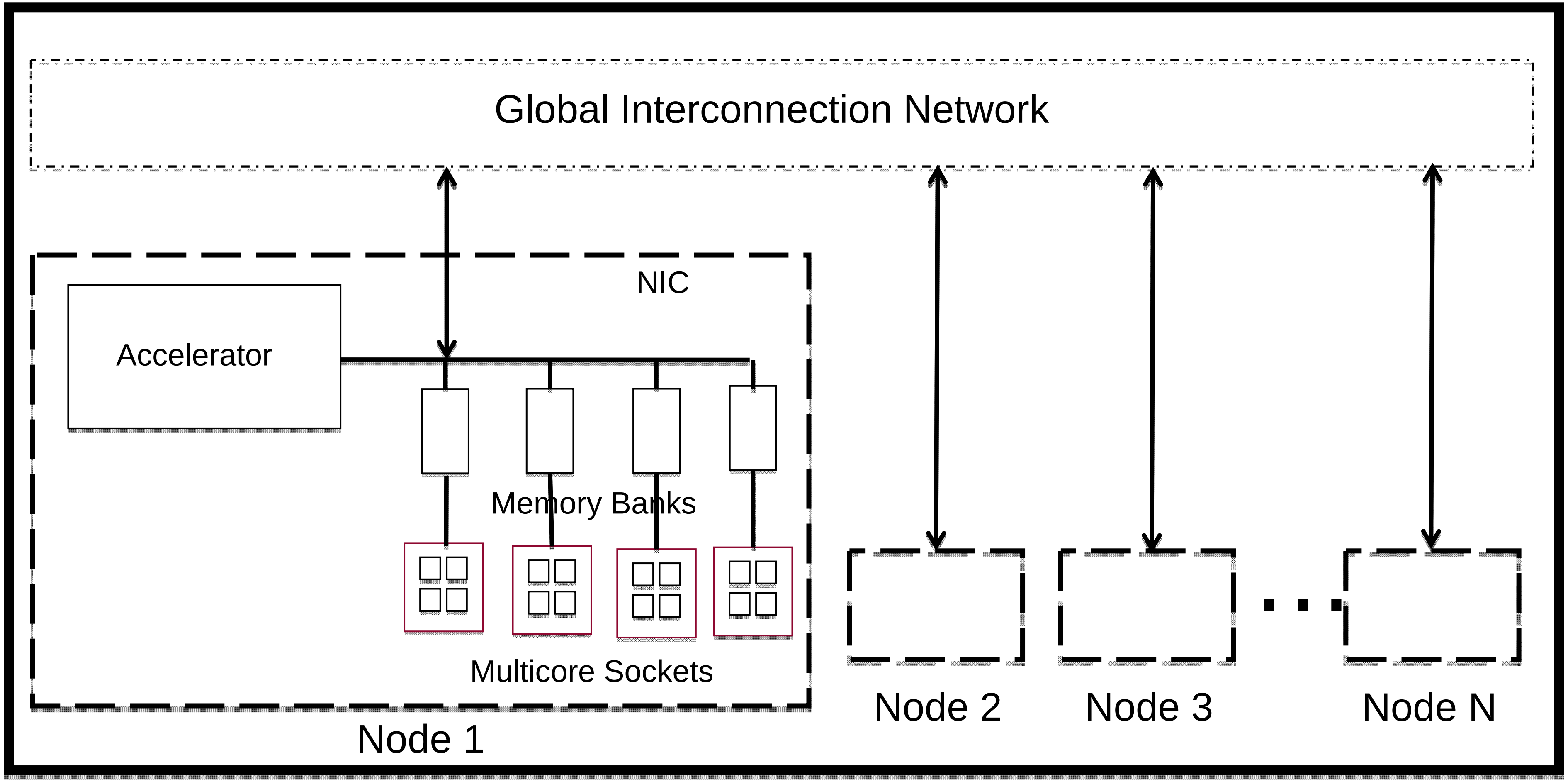




Assumptions of Conventional Practice

- Binary bit data hardware representation
  - Base-2 logic and storage
- Single word actions format
  - Floats, ints, Booles, pointers
- Program counter control state
  - serialization
  - Plus stack-frame pointers
- Distributed memory, BSP, SPMD, message-passing
- Separation of processing logic and memory storage
  - Von Neumann bottleneck
- Fastest nodes possible with fastest sockets with fastest cores
- Moore's Law and Dennard scaling
  - Dying and dead
  - Semiconductor technology and dies, with some optics
- Static compile/load-time resource allocation and task scheduling
- Legacy codes and programming interface with incremental changes

Advanced GAS Exascale System Architecture



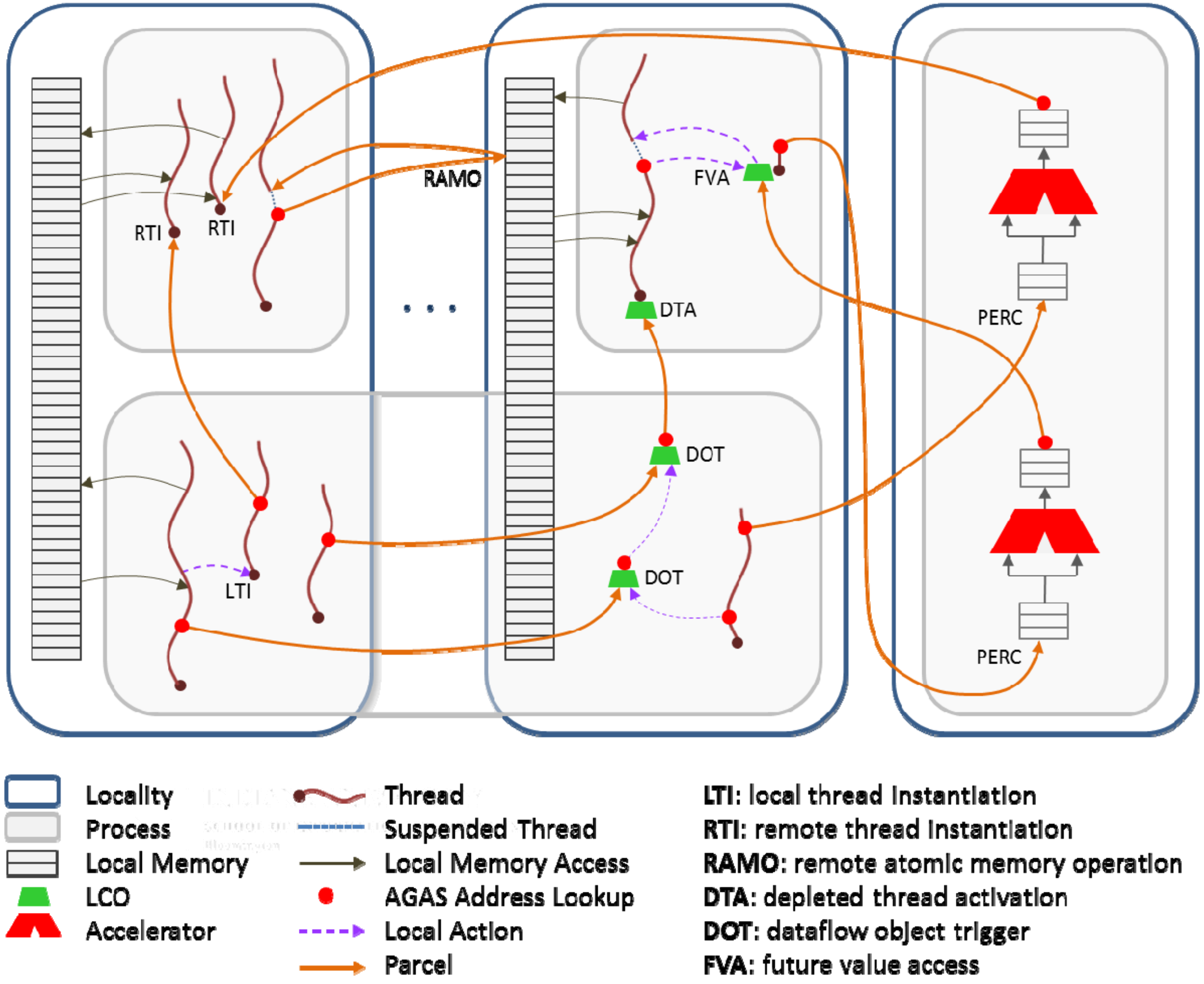
Performance Factors - SLOWER

$$P = e(L,O,W) * s(S) * a(R) * U(E)$$

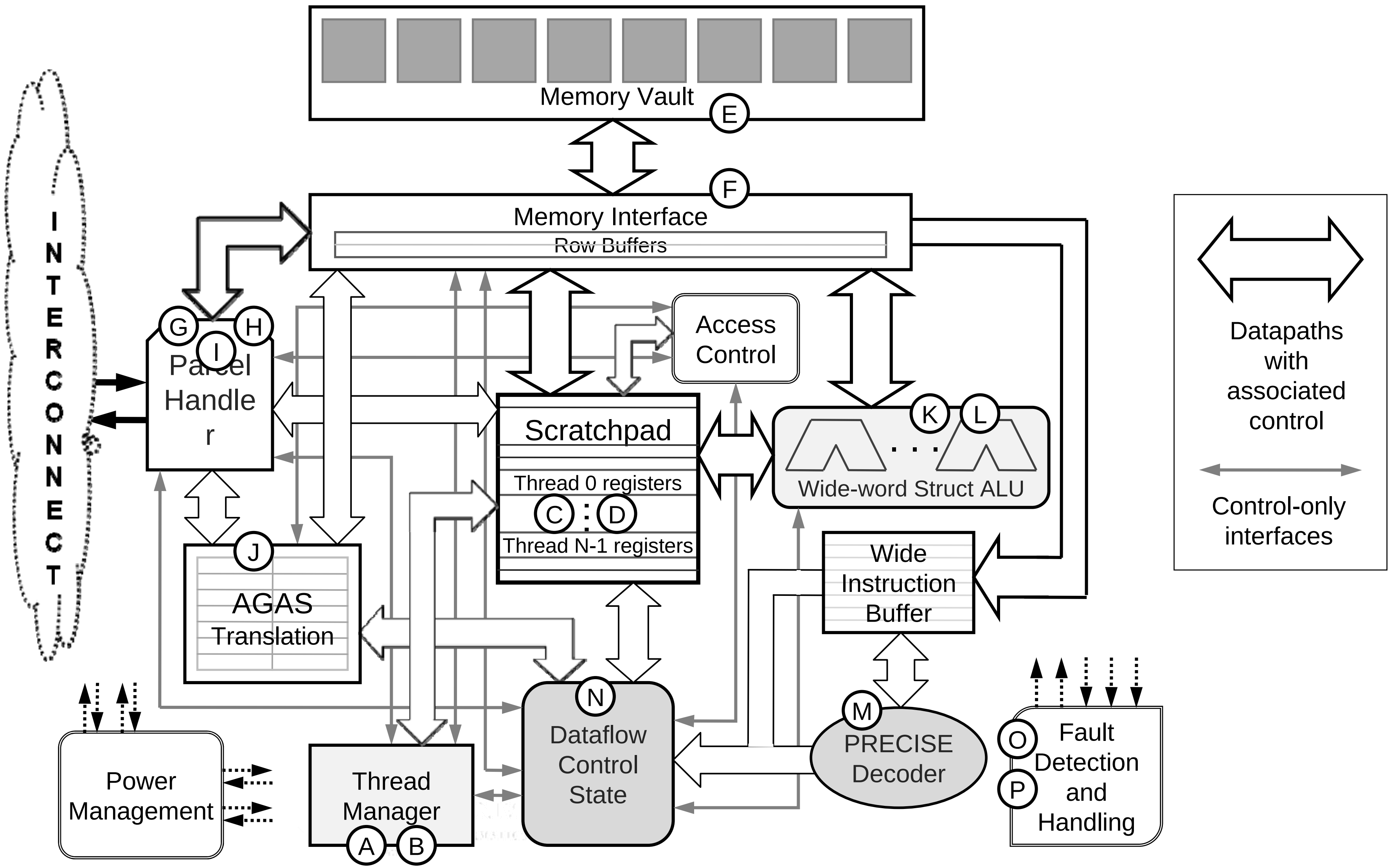
P – average performance (ops)
e – efficiency (0 < e < 1)
s – application's average parallelism,
a – availability (0 < a < 1)
U – normalization factor/compute unit
E – watts per average compute unit
R – reliability (0 < R < 1)

- Starvation
  - Insufficiency of concurrency of work
  - Impacts scalability and latency hiding
  - Effects programmability
- Latency
  - Time measured distance for remote access and services
  - Impacts efficiency
- Overhead
  - Critical time additional work to manage tasks & resources
  - Impacts efficiency and granularity for scalability
- Waiting for contention resolution
  - Delays due to simultaneous access requests to shared physical or logical resources

Semantic Components of ParalleX



LTI: local thread instantiation
RTI: remote thread instantiation
RAMO: remote atomic memory operation
DTA: depleted thread activation
DOT: dataflow object trigger
FVA: future value access
PERC: parcelation

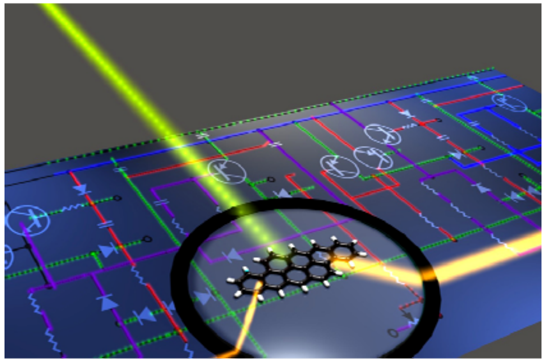

Memory Vault
E
F
Memory Interface
Row Buffers
INTERCONNECT
G
H
I
Parcel Handler
Access Control
Scratchpad
Thread 0 registers
C
D
Thread N-1 registers
K
L
Wide-word Struct ALU
J
AGAS Translation
Wide Instruction Buffer
N
Dataflow Control State
M
PRECISE Decoder
O
P
Fault Detection and Handling
Power Management
Thread Manager
A
B
Datapaths with associated control
Control-only interfaces

Neo-Digital Age

- Goal and Objectives
  - Devise means of exploiting nano-scale semiconductor technologies at end of Moore's Law to leverage fabrication facilities investments
  - Create scalable structures and semantics capable of optimal performance (time to solution) within technology and power limitations
- Technical Strategy
  1. Liberate parallel computer architecture from von Neumann (vN) archaism for efficiency and scalability; eliminate vN bottleneck
  2. Rebalance and integration of functional elements for data movement, operations, storage, control to minimize time & energy
  3. Emphasize tight coupled logic locality for low time & energy
  4. Dynamic adaptive localized control to address asynchrony and expose parallelism with emerging behavior of global computing
  5. Innovation in execution model for governing principles at all levels

Neo-Digital Age – extending past foundations

- Near nano-scale semiconductor technology
  - Flat-lining of Moore's Law at single-digit nano-meters
  - Cost benefits of fab lines for economy of scale through mass-market
- Logic function modules
  - Exploit existing and new IP for efficient functional units
  - ALUs, latches/registers, nearest-neighbor data paths
- Integrated optics
  - Orders of magnitude bandwidth increase
  - Inter-socket
  - On chip



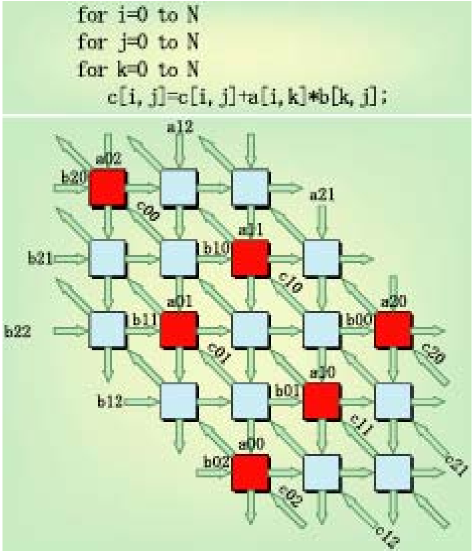
- Advanced packaging and cooling
  - Dramatic improvement opportunities in volumetric utilization
  - 3-D integration of combined memory/logic/communication dies
  - Return to wafer-scale integration

Neo-Digital Age – Principles

- Elimination of vN-based parallel architecture
  - Constrained parallel control state
  - Processor-centric approach optimizes for ALU utilization with very deep memory hierarchy; wrong answer
- ALU-pervasive structures
  - Merge ALUs with storage and communication structures
  - High availability of ALUs rather than high utilization
  - Slashes access latencies to save time and energy
- Cells of logic/memory/data-passing for single-cycle actions
  - Optimized for space/energy/performance
  - Optimized for memory bandwidth utilization
- Emphasis on fine-grain nearest-neighbor data-movement structures
  - Direct access to adjacent state storage
  - Enables communication through nearest neighbor
- Virtual objects in global name space
  - Data, instructions, synchronization
  - Intra-medium packet-switched wormhole routing
  - Dynamic allocation

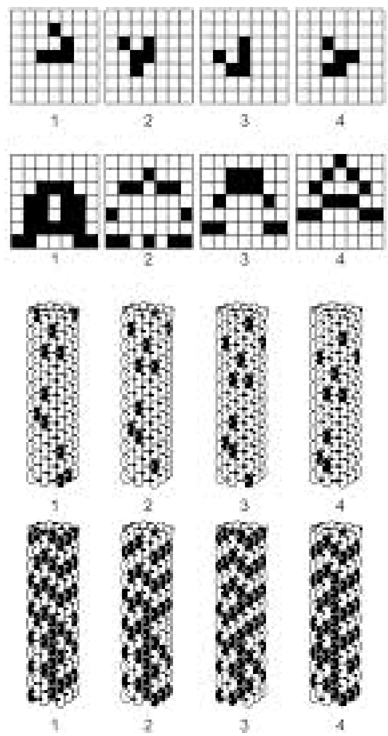
Prior Art: Non-von Neumann Architectures

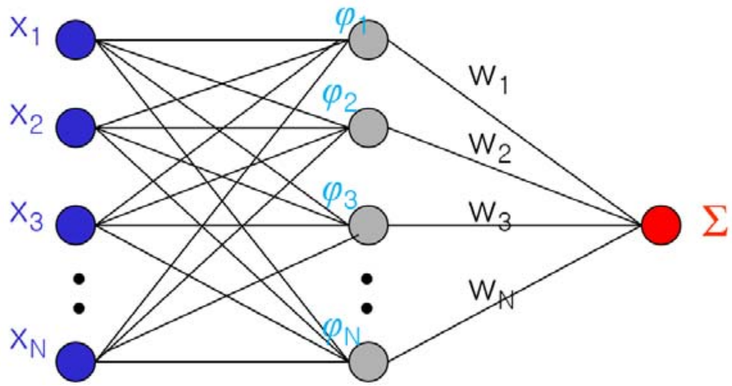
- Dataflow
  - Static (Dennis)
  - Dynamic (Arvind)
- Systolic Arrays
  - Streaming (HT Kung)
- Neural Networks – connectionist
- Processor in Memory (PIM)
  - Bit or word-level logic on-chip at/near sense amps of memory
  - SIMD (Iobst) or MIMD (Kogge)
- Cellular Automata (CA)
  - Global emergent behavior from local rules and state
  - von Neumann (ironically)
- Continuum Computer Architecture
  - CA with global naming and active packets (Sterling/Brodowicz)

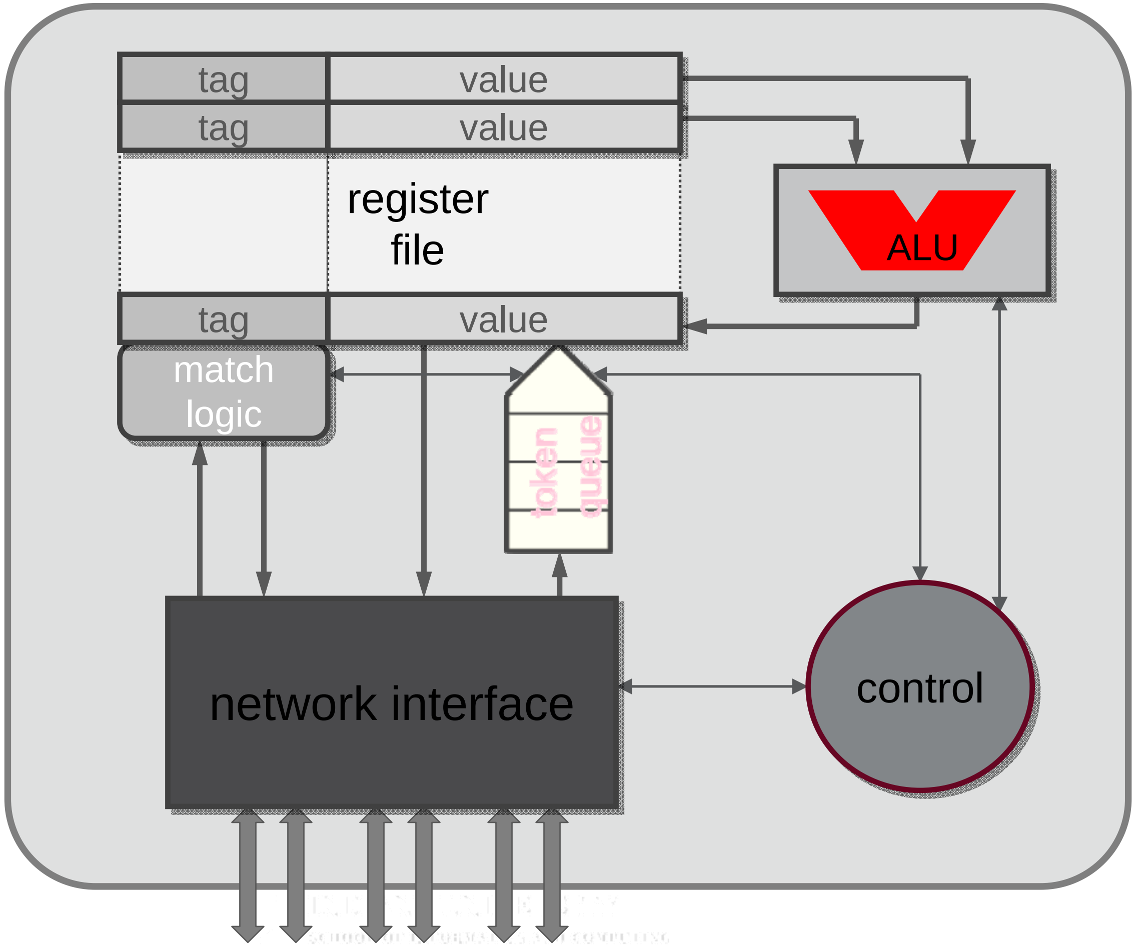


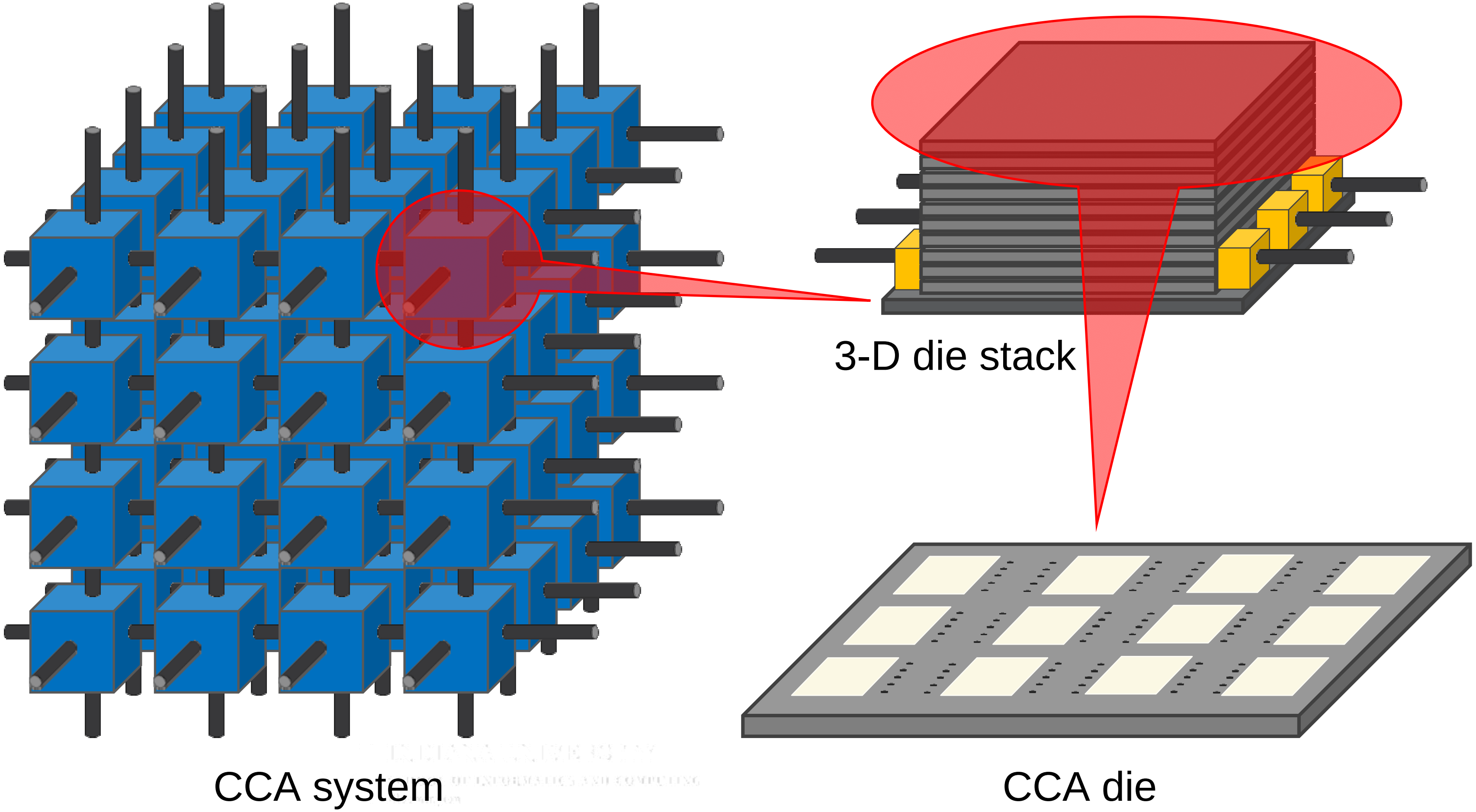
Properties

- Local communication
  - Systolic array: mostly
  - Dataflow: classically not
  - Processor in Memory: logic at the sense amps
  - Cellular Automata: fully
  - Continuum Computer Architecture: fully, with pipelined packet switche
  - Neural networks: not
- Event driven for asynchrony management
  - Systolic array: classically not, iWarp yes
  - Dataflow: yes
  - PIM: no
  - Cellular Automata: can be
  - CCA: yes
  - Neural networks: classically no, neuromorphic yes
- Merged logic/storage/communication
  - Systolic array: yes
  - Dataflow: classically not, possible
  - PIM: yes
  - Cellular Automata: yes
  - CCA: yes
  - Neural networks: yes






tag
value
tag
value
register
file
tag
value
match
logic
ALU
token
queue
network interface
control


3-D die stack
CCA system
CCA die

Performance

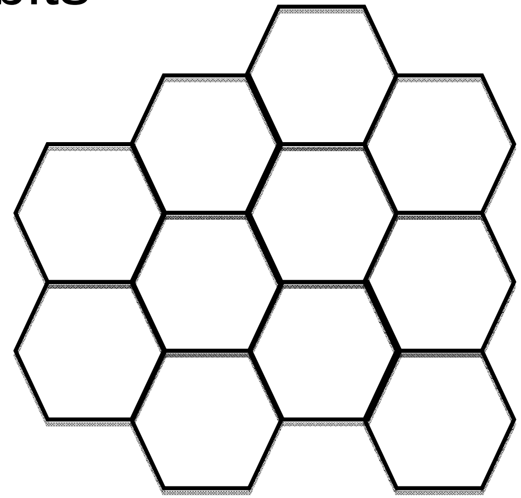
- Maximum ALUs
  - For high utilization
- Maximum memory bandwidth
- Lots of inter-cell communication bandwidth
- Reduced overhead
- Reduced latency
- Adaptive routing for contention avoidance
- Multi-variate storage beyond the bit
- Multi-variate logic beyond base-2 Boolean
- Dynamic adaptive execution model

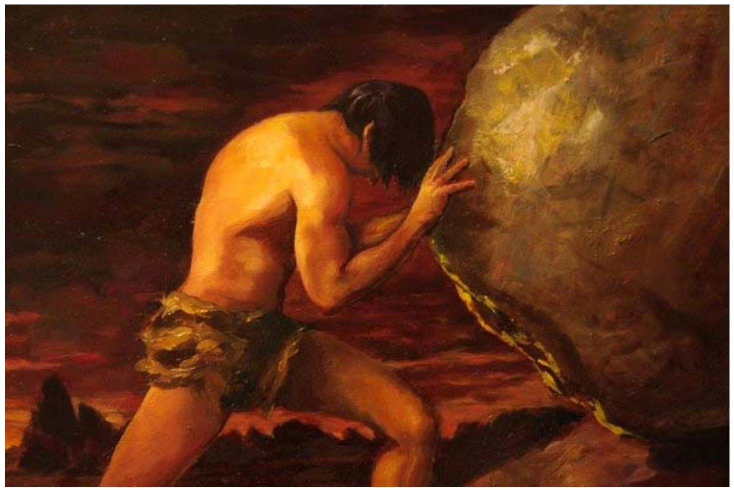
Energy

- Minimize distance between elements
  - Permeate structures with ALUs
  - Short latencies between memory & registers
- Multiple clock rates to match timings between logic and storage and
- Neighbor cell memory/register access
- Eliminate large caches and multi-level caches
- Eliminate speculative execution

Questions can be Answered Now

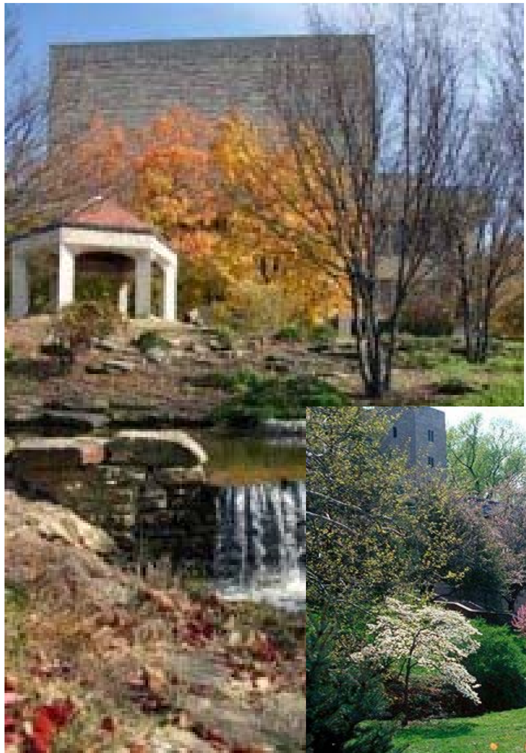
1. Trade-offs of DRAM bits and SRAM bits
   – Space, Time, Energy
2. Cell rules
   – derived from parallel execution model
3. Cell granularity
   – Breakpoint where mitosis is better
4. Design of logic cell (Fonton)
   – Very simple compared to full conventional processors
5. Reference implementation
   – Emulation, Simulation, FPGA, ASIC, custom
6. 3-D packaging
7. Software environment
8. Application analysis





Conclusions

- Moore's Law is flat-lining
- Architecture may deliver another performance boost
- Neo-digital age defined as post von Neumann architecture era using advanced semiconductor technology
- Exploitation of nearest neighbor structures, ALU intensive for low latency, lightweight cells for high concurrency
- Requires alternative execution model
- All R&D can be undertaken now




CardCow.com