JUMP

Joint University Microelectronics Program

Proposers' Day Workshop

Monday, January 23, 2017

@srcJUMP, #JUMPpdw

JUMP

Joint University Microelectronics Program

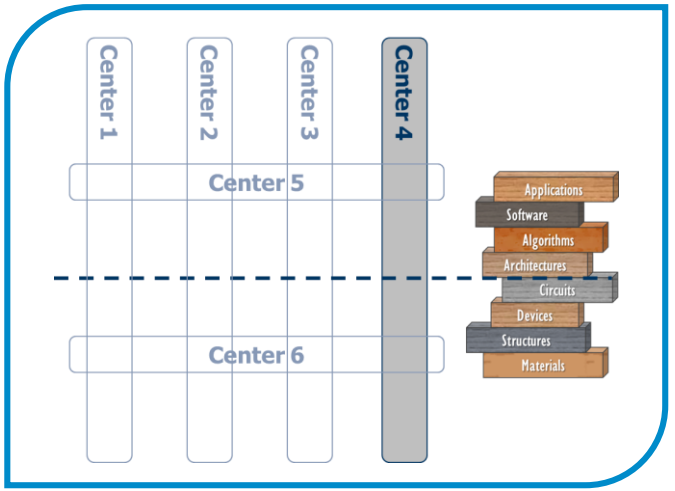
Intelligent Memory and Storage

Sean Eilert

Fellow

Micron Technology

Vertical Research Center



JUMP

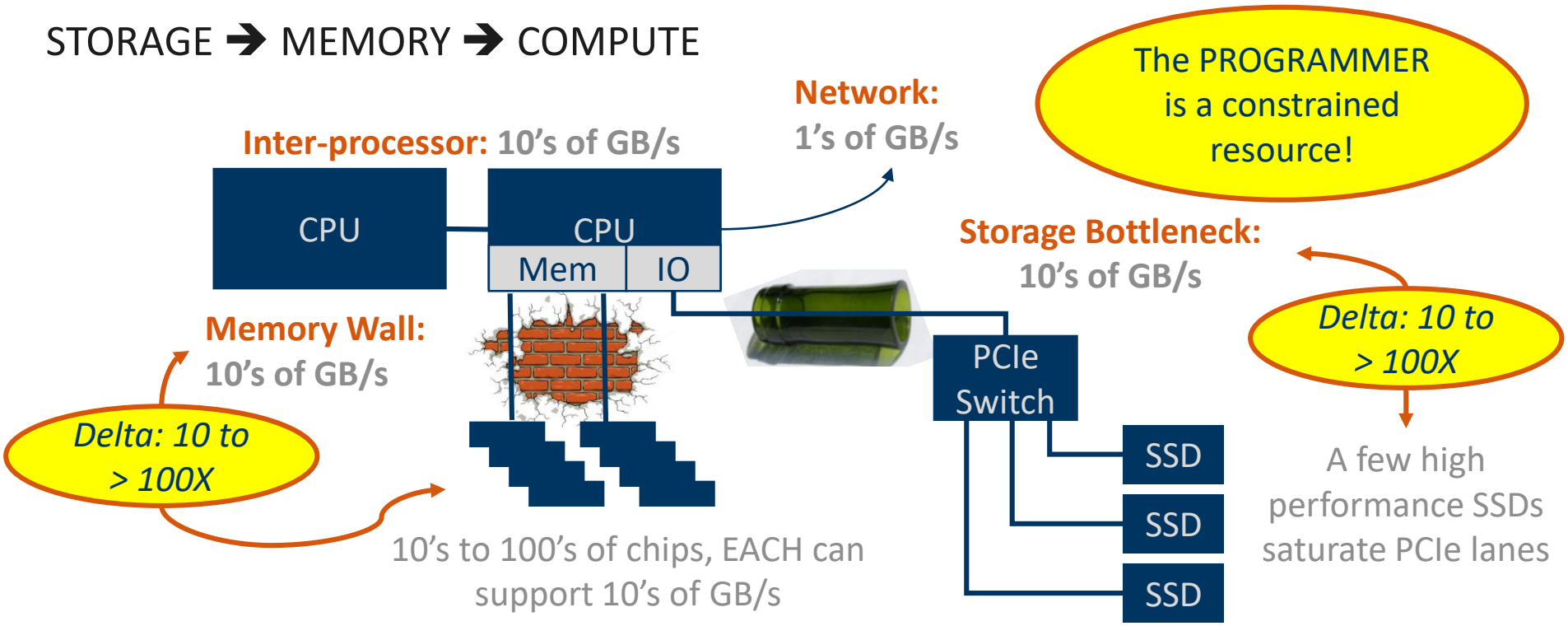
Joint University Microelectronics Program

High Level Overview

Conventional Bottlenecks

TODAY'S HIGH PERFORMANCE LOGIC IS OFTEN LIMITED BY DATA MOVEMENT

STORAGE ➔ MEMORY ➔ COMPUTE



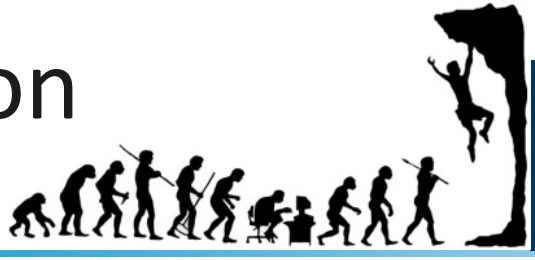
Example Capacity Optimized Platform:

1536 Memory Dice
(2 sockets*4 chan/socket*3 dimms/chan *4 ranks/dimm*16 die/rank)
Measured Bandwidth=109GB/s (72MB/s/die)

Memory IO bandwidth (2400MT/s) = 2.4GB/s/die
Memory Bank Bandwidth (1KB/45ns) = 22GB/s

30-300X Opportunity

Evolution



COTS Scale Out

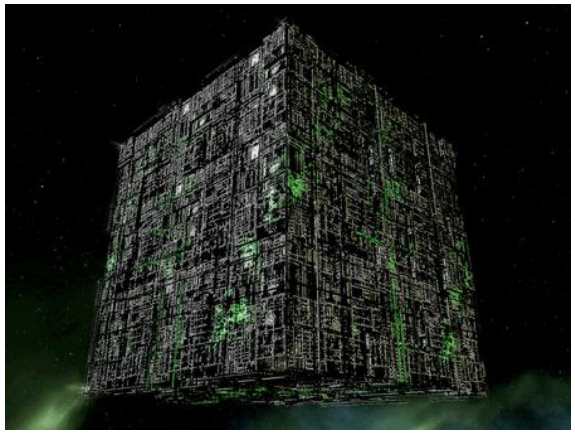
- Applications parallelized onto commodity hardware
- Limited interconnect between nodes

Custom Silicon DataCenters

- High Dimensional Network
- Tightly coupled Processor, Memory, and Storage
- In-node accelerators

Intelligent Memory & Storage

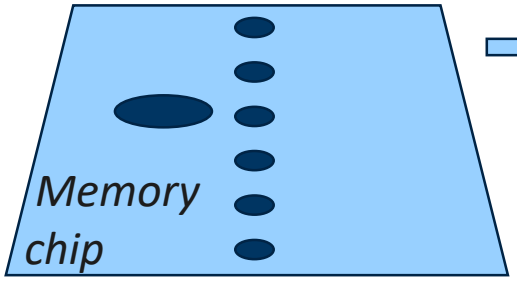
- Efficient Programming Model
- High Dimensional Fabric of Intelligent memory and storage
- Enormous cross-sectional bandwidth and memory density
- Fabric-wide Operating System
  - Optimization based on System-wide preferences and dynamic state



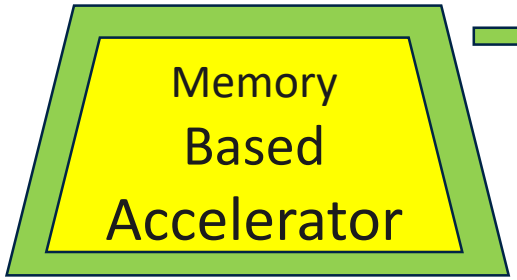
Example Bandwidth optimized architecture

General purpose programmable logic for filtering, aggregation, first-pass calculation:
For Example:

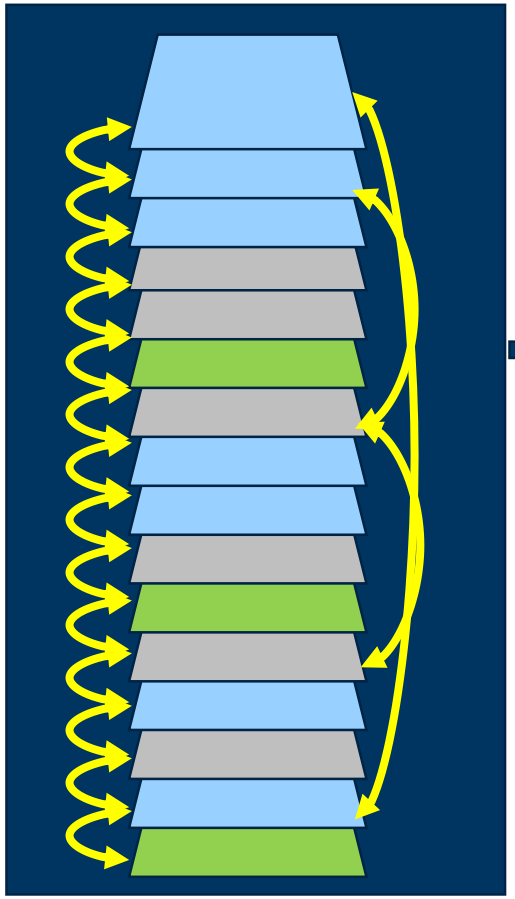
1 x RISC-V Rocket @ 200MHz
+
10 x RISC-V Z-SCALE @ 100MHz



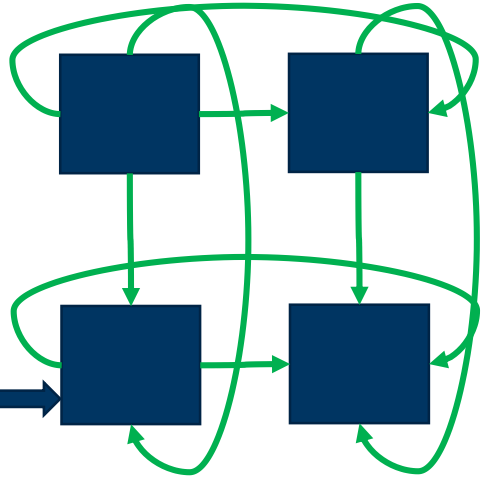
+ Special purpose accelerators, perhaps based on memory technologies



Dense Packaging and/or 3D Integration
With high bandwidth, low energy internal networks



Dense Board design with inter-package fabric



- High Bandwidth links to extend networks between enclosures
- CPU's, GPUs included in enclosures as-needed

Potential & Challenges

Hardware

- 1-2u Enclosure
- ~100K Silicon Components
- ~1M small processing elements
  - ~1 processor / GB
- 3KW power budget
- Internal thermal constraints
- Large Bisection Bandwidth (depends on network)
- ~100TB/s total die-die bandwidth
- >1PB/s total die-local bandwidth
- <100GB/s external bandwidth

Runtime System Management Challenges

- Programming model abstracting both ALGORITHMS and HARDWARE
- Optimization engine to map algorithms onto available hardware given the **current state** of the system
- Dynamic Algorithm Optimization across constrained resources
  - Local vs Global
  - Thermal
  - Power
  - Memory Allocation and consistency
  - Authentication
  - Replication for performance or resilience
  - Management of internal failures

Call To Action

Establish Relevant Metrics
E.g. (Correct Decision Rate or latency) per (Silicon area, System Volume, or Watt)

Research fundamental parallelization limits of Algorithms and Applications
Establish abstraction languages for algorithms and accelerators

Prototype run-time system to optimally map algorithms to accelerators based on system configuration, programmer preferences, and current dynamic state

Reference Platform
Deliver run-time software framework
Establish hardware sufficient to evaluate applications and run-time system at scale

Understand bounds of possible hardware implementations
Explore effectiveness of memory based accelerators and dense processing technologies
Understand and leverage parallels between in-memory compute & in-sensor processing

JUMP

Joint University Microelectronics Program

Detailed Overview

Center Motivation

- Advances in information technology have pushed data generation rates and quantities to a point where memory and storage are the **focal point of optimization** of computer systems.
- Transfer **energy, latency and bandwidth** are critical to performance and energy efficiency of these systems.
- The solutions to many modern computing problems involve **many-many relationships** that can benefit from high cross-sectional bandwidth of the distributed computing platform.
- As an example, large scale graph analytics involve high cross-data-set evaluation of numerous neighbor relationships ultimately demanding the highest possible cross-sectional bandwidth of the system.

Research Considerations (1)

- This research vector seeks a **holistic, vertically-integrated, approach** to high-performance Intelligent Storage systems encompassing the operating system, programming models, memory management technologies, and a prototype system architecture.
  - Significant advancements may be made by utilizing distributed compute elements and accelerators **in close physical proximity** to the location of data storage.
  - An optimal hardware platform may contain a large fabric of interconnected memory and storage devices with large cross-sectional bandwidth, integrated logic and accelerators in conjunction with islands of conventional high performance computing elements.
  - **Gains of orders of magnitude** in volumetric density, information processing density, power or performance should be pursued.
- A primary focus area for this center will be in establishing an operating system framework allowing **run-time optimization of the system** based on system configuration preferences, programmer preferences, and the current state of the system.
  - There will be many run-time optimization challenges in such a heterogeneous platform containing Near-Memory Von Neumann and non-Von Neumann elements as there are inherently numerous means to achieve the same end (e.g. compute slowly locally, transfer data and compute quickly remotely, compute very efficiently but approximately in an accelerator) and proper optimization may differ depending on the current state of the system.

Research Considerations (2)

- As an envisioned system example, execution time and data movement are automatically balanced in a near-optimal fashion on simple hardware through OS and hardware hooks without the requirement of, but with the allowance of, programmer intervention; where built-in controls allow a system administrator to tradeoff performance, power efficiency, response latency, etc. to operate within the constraints of the specific system installation.
- Scheduling decisions may involve proper use of local and global metrics of bandwidth, power, latency, and temperature, etc.

- In order to reasonably establish run-time optimization algorithms, **suitable metrics for performance of such a system must be established** to measure information processing density. e.g. ("decision rate" or "correct decision rate") per ("Kg of silicon", "cm^3", or "Watt").
  - For a simple example, imagine a 1u enclosure containing > 100K memory dice interconnected in a high dimensional fabric, each capable of accessing and operating on data locally and each consuming 1 watt of power in its most active state.
  - Algorithms for power throttling based on temperature and power consumption would be required for this system and these algorithms likely involve local and global aspects of the control system.
  - System theoretic bounds on information processing density should be established to demonstrate the research areas of highest potential.
  - It is foreseen that replication will be a required aspect of this type of system for performance and for resilience and that this system should be designed to co-optimize this replication with retention and endurance management and authentication strategies (which users are allowed to run which programs on which data).

Theme Requires

- This holistic approach must keep in focus the **relevant, emerging memory** technologies and their respective, novel system architectures and hierarchies (including subsystems and their caches), as well as the advanced materials and processes required to manufacture them.
- An eye towards **backward compatibility** where possible is also of interest to facilitate migration of applications to this new framework.
- Proposers are expected to define a **grand challenge** in the Intelligent Memory and Storage space that their Center will address.

Possible Research Tasks (1)

In addition to the broad theme described above, possible research tasks of interest includes but are not limited to:

Compute in memory systems

- Memory at a higher level
- Compute in Memory – Von Neumann (e.g. PIM)
- Compute in Memory – Non Von Neumann (e.g. Brain-Inspired)
- Heterogeneous Systems
  - Systems on chip with novel memory hierarchy
  - Nonvolatile computing (HW and SW)
  - In-memory computing
  - Reconfigurable computing
- Inference kernels
  - Learning algorithms at a low level
  - Sensitivity analysis on how process variation might affect accelerators

New Architecture and Programming paradigms, Self-optimizing Systems Allowing for Appropriate Programmer Control

- New programming and architecture paradigms to exploit non-volatile and low latency emerging memory systems

Possible Research Tasks (2)

In addition to the broad theme described above, possible research tasks of interest includes but are not limited to:

New Architecture and Programming paradigms, Self-optimizing Systems Allowing for Appropriate Programmer Control (cont.)

- Algorithms and memory architectures that facilitate decision making based on approximate, or error-prone, stored data
- Address the myriad of architecture, software, and applications implications of new memory devices and alternative memory hierarchies over the cache hierarchy, system memory, and storage tiers
- Scalable, reconfigurable, energy efficient information extraction (circuits/ systems/ algorithms/ architecture). Includes Distributed, cognitive, non-traditional computing and analog
- 10X more power efficient computing platform scalable from high performance application processors to less-demanding processors for IoT/sensors/etc. with cost awareness. The technology can span across material, devices, packaging, circuits/systems techniques, computer architecture including but not limited to heterogeneous computing, memory technology (including NVM) and high-speed interface (on-chip and off-chip), etc.
- Heterogeneous Systems - Computing on non-traditional fabrics (e.g. cross points, FPGA as a compute unit. etc.)

Possible Research Tasks (3)

In addition to the broad theme described above, possible research tasks of interest includes but are not limited to:

New Architecture and Programming paradigms, Self-optimizing Systems Allowing for Appropriate Programmer Control (cont.)

- Co-designed HW and algorithms
  - Computational theory for Neuromorphic
  - Computational complexity (time and memory) of non von Neumann
  - Computational Complexity (time and memory) of neuromorphic
  - Computational Complexity (time and memory) of stochastic and random computing
  - Scalability of Neuromorphic and Stochastic computation

Small, Probably Low Cost, Compute+Memory+Sensor Node Capable of making Basic Decisions/observations and Reporting to a Larger System

- Small scale localized memory and storage systems (e.g. IOT)
- In-sensor memory and compute (for energy efficient sensors systems)

Intersection of Big-data Analytics and Big-data for Sensor Data/Lazy Computation of Sensor Data Local to Data Storage

- Architectures to optimize data movement and memory-compute partitioning for energy and performance

Possible Research Tasks (4)

In addition to the broad theme described above, possible research tasks of interest includes but are not limited to:

Intersection of Big-data Analytics and Big-data for Sensor Data/Lazy Computation of Sensor Data Local to Data Storage (cont.)

- Architectures to optimize data movement and memory-compute partitioning for energy and performance
- Big data, large-scale, distributed memory and storage systems (e.g. Data Center)
- Heterogeneous big data and distributed analytics
- Intersection of Big-Data Analytics and Big-Data for sensor data / lazy computation of sensor data local to data storage

New Technologies, Materials and Processes

- Identification and development of new memory-device technologies
  - Highly-scalable, non-volatile memory exhibiting high durability with low variation/stochasticity
    - Low latency, energy-efficient
  - Non-volatile memories suitable for embedded implementations including low-power computing and SOC
  - Ultra-low power, moderate-density, integrated memory for sensors
  - Access devices for cross-point and 3D geometries

Possible Research Tasks (5)

In addition to the broad theme described above, possible research tasks of interest includes but are not limited to:

New Technologies, Materials and Processes (cont.)

- Advanced charge-based and steep-slope devices.
  - Phase change materials
  - Metal insulator transition, mott insulators
  - 2D materials (e.g. TFET)
  - New ferro-electrics (highlighting reliability)
- Materials and interfaces for spin-based logic and memory
  - High polarization magnets/interfaces for spin injection
  - High polarization and low coercivity ferroelectrics
  - Magnetoelectrics, multiferroics, and uniferroics with ME
  - Spin-orbit coupled materials
  - Topological materials

Possible Research Tasks (6)

In addition to the broad theme described above, possible research tasks of interest includes but are not limited to:

New Technologies, Materials and Processes (cont.)

- Novel processing for advanced devices and scaled geometries (incl. ALD and Atomic Layer Etching)
- Novel materials and means for reducing processing temperatures, especially B/E processes.
- Advanced materials and processes for self-assembly.
- Novel organic electronic materials (semiconductor & conductor)
- **Functional 3D integration (monolithic 3D ICs)**
- Holistic approach to harnessing advanced memory/storage devices and complimentary CMOS (and beyond) with supporting interconnect and packaging technologies targeted and optimized for non-traditional and cognitive computing systems.
- Demonstration of novel, vertical integration and leveraging of advanced CMOS and beyond, memory/storage, and interconnect device technologies, to achieve dramatically improved energy/performance versus state-of-the-art CMOS. Demonstrate these techniques on realistic computational problems by prototyping functional units/cores/accelerators.

Possible Research Tasks (7)

In addition to the broad theme described above, possible research tasks of interest includes but are not limited to:

Authentication & Resilience (cont.)

- Memory management and control technologies including error management and authentication
- **Resilient/self-healing compute and memories**
- Disruptive memory architectures and technologies to optimize intelligent and flexible storage design, management, processing and system scalability to mitigate data explosion challenges with radiation hardening capability.

IO & Networking

- In-Memory networking structures
- Virtual=physical large memories (no translation Software)

Intelligent Memory and Storage

NEW TECHNOLOGIES, MATERIALS, AND PROCESSES

****All material and device efforts should be coupled closely with novel techniques for advanced modeling including atomistic first principles and electron transport**

New memory-device technologies

- Highly-scalable, non-volatile memory exhibiting high durability with low variation/stochasticity
  - Low latency
  - Energy-efficient
- Non-volatile memories suitable for embedded implementations including low-power computing and SOC
- Ultra-low power, moderate-density, integrated memory for sensors
- **Access devices for cross-point and 3D geometries**

- **Advanced charge-based and steep-slope devices**.
  - Phase change materials
  - Metal insulator transition, mott insulators
  - **2D materials (e.g. TFT)**
  - New ferro-electrics (highlighting reliability)
- **Materials and interfaces for spin-based logic and memory**
  - High polarization magnets/interfaces for spin injection
  - High polarization and low coercivity ferroelectrics
  - Magnetoelectrics, multiferroics, and uniferroics with ME
  - Spin-orbit coupled materials
  - Topological materials
  - Physic of critical oxide technologies (e.g. MgO, MgAlO, BiFeO) including fatigue and failure

Intelligent Memory and Storage

NEW TECHNOLOGIES, MATERIALS, AND PROCESSES

****All material and device efforts should be coupled closely with novel techniques for advanced modeling including atomistic first principles and electron transport**

New memory-device technologies

- **Novel processing for advanced devices and scaled geometries (esp. ALD and Atomic Layer Etching)**
- Novel materials and means for reducing processing temperatures, especially B/E processes.
- Advanced materials and processes for self-assembly.
- Functional 3D integration (monolithic 3D ICs)
- Novel organic electronic materials (semiconductor & conductor)

Ultimate goals

- Holistic approach to harnessing advanced memory/storage devices and complimentary CMOS (and beyond) with supporting interconnect and packaging technologies targeted and optimized for non-traditional and cognitive computing systems.
- Demonstration of novel, vertical integration of advanced **CMOS and beyond**, memory/storage, and interconnect technologies, realizing dramatically improved energy/performance versus state-of-the-art CMOS. functional units/cores/accelerators.

Degrees of Emphasis

While Centers are not limited to students from these disciplines, **students obtaining degrees** from the Intelligent Memory and Storage Center in the following areas of study are of particular interest to our sponsors in the years ahead:

MS / PhD

- **Applied Math**
- **Applied Physics, Physics**
- **Bioengineering**
- **Chemistry or Chem Eng.** (growth, synthesis, fabrication processes)
- **Comp Sci. & Comp. Eng.**
- **Electrical / Electronics Engineering or Electrical and Computer Engineering**
- **Material Science or Engineering**
- **Systems Engineering**

JUMP

Joint University Microelectronics Program

@srcJUMP, #JUMPpdw