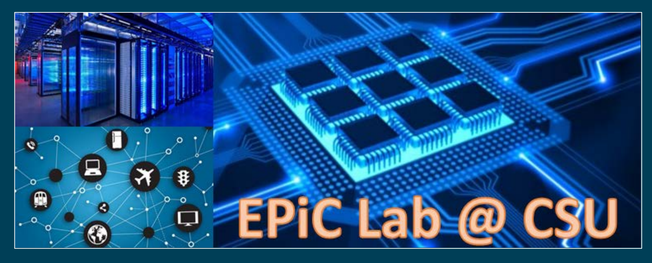
Embedded Systems and High Performance Computing (EPiC) Lab

Research Sponsors

Prof. Sudeep Pasricha (Director)
Monfort and Rockwell-Anderson Professor
Dept. of Electrical and Computer Engineering | Dept. of Computer Science
Colorado State University, Fort Collins, CO – 80523
email: sudeep@colostate.edu

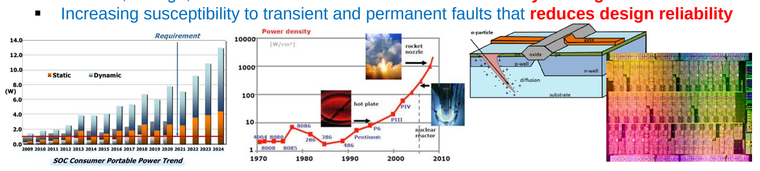


Mission: Algorithms and architectures for energy-efficient, fault-tolerant, and secure design of **embedded systems** (cyber-physical systems), **mobile computing** (smartphones, wearables, internet-of-things), and **high performance computing** (datacenters, supercomputers)

CAD Tools for Multicore Chip Design

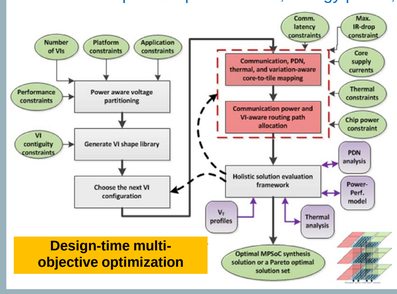
- **Nearly all modern innovations depend on continued advances in multicore system-on-chip computing performance**
  - **Major impact on innovation across application domains: automotive, defense, medical, multimedia, telecommunications, aerospace, mobile/cloud computing**

- **But multicore system-on-chip design in advanced semiconductor fabrication technologies today faces several challenges**
  - High power/energy dissipation that **increases costs** and **limits achievable performance**
  - Process, voltage, and thermal variations that cause **uncertainty** and **high time-to-market**
  - Increasing susceptibility to transient and permanent faults that **reduces design reliability**

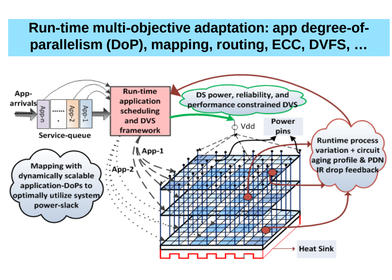


Need new computer-aided design (CAD) tools to perform multi-objective chip design exploration and optimization

- **Novel CAD tools for emerging 2D/3D multicore chip design**
  - Design-time algorithms for core/memory/network selection and configuration
  - Run-time algorithms to map computation, communication, and data on the chip die
  - Co-optimize: performance, energy/power, soft/hard fault resilience, security, yield, cost, …



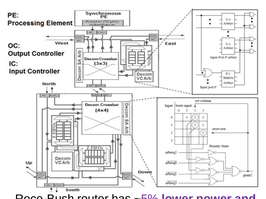
Run-time multi-objective adaptation: app degree-of-parallelism (DoP), mapping, routing, ECC, DVFS, …

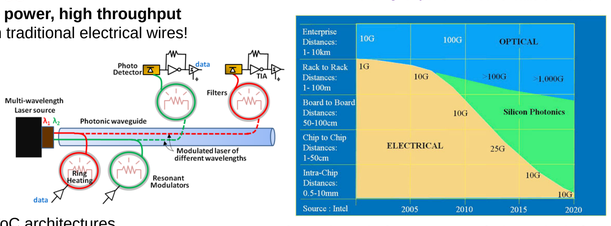


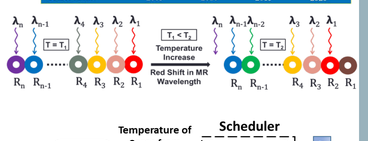
Network-on-Chip (NoC) Architectures

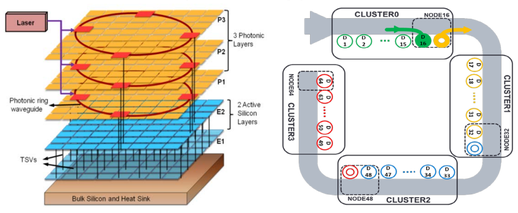
Design of on-chip communication fabric is a very critical factor influencing multicore chip performance, power, and reliability

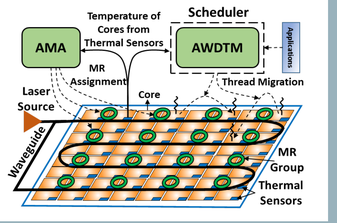
- **NoCs have replaced on-chip buses, but face challenges**
  - High packet transfer latency with increasing core counts
  - High susceptibility to transient (soft) and permanent (aging/hard) faults
  - Need to balance multiple goals while satisfying design constraints
- **Fault-tolerant NoC protocols and adaptation**
  - **Reliable NoC packet routing algorithms**
    - OE+IOE: hybrid multiple turn-model routing algorithm for 2D NoCs
    - 4NP-First: hybrid turn-model routing algorithm for 3D NoCs
  - **Fault vulnerability aware NoC optimizations**
    - Proposed network vulnerability factor (NVF) metric to characterize vulnerability of network interfaces and routers to faults
- **NoC architecture optimization**
  - **Roce-Bush NoC router for 3D NoCs**
    - Routing algorithm-aware decomposition of NoC router
  - **Memory- and application-aware NoC prioritization**
    - Heterogeneous NoC scheduling with anti-starvation support
- **Photonic NoC architectures**
  - **Speed of light latency, low power, high throughput**
    - Scale much better than traditional electrical wires!
  - **Challenges:**
    - Photonic crosstalk
    - Process variations
    - Thermal variations
    - Topology design
    - Protocol design
    - Fabrication choices
    - Off-chip interfacing
  - **Solutions:**
    - New 2D/3D photonic NoC architectures
      - Free-space, mono-layer, multi-layer
    - New cross-layer (device, circuit, system) techniques
      - To overcome crosstalk and variation uncertainty
    - New photonic arbitration, encoding, flow-control protocols
    - New off-chip CPU-memory photonic architectures





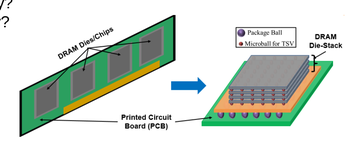


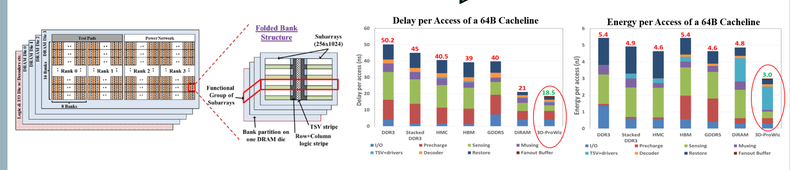




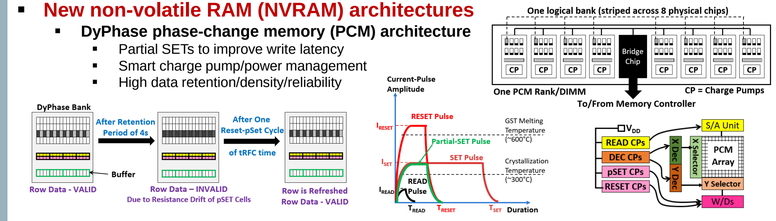
Memory Architectures

- **Improvement in memory density, bandwidth, and form factor are critical for next generation multicore computing chips**
  - To support increasing data demands from rising chip core counts, growing graphics capabilities, the Big Data revolution, and higher network/Ethernet speeds
  - **Challenges:**
    - How to scale memory component density?
    - How to increase bandwidth and reduce latency?
    - How to best manage power dissipation/energy?
- **New 3D DRAM architectures**
  - **Decomposed (folded) bank architecture**
    - Split bank and rank across layers (3D-ProWiz)
    - Improved performance over state-of-the-art DRAMs
      - HMC, HBM, DDR/GDDR, DiRAM, …

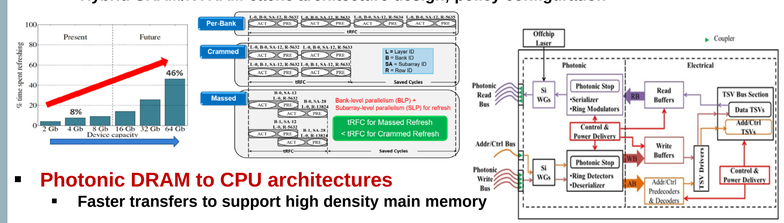




- **New non-volatile RAM (NVRAM) architectures**
  - **DyPhase phase-change memory (PCM) architecture**
    - Partial SETs to improve write latency
    - Smart charge pump/power management
    - High data retention/density/reliability



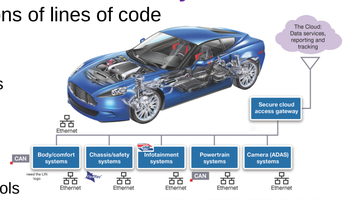
- **DRAM and cache/scratchpad memory optimizations**
  - **DRAM refresh overhead reduction (new massed/crammed refresh techniques)**
  - **Scratchpad data placement (static and adaptive packing strategies)**
  - **Hybrid SRAM/NVRAM cache architecture design; policy configuration**



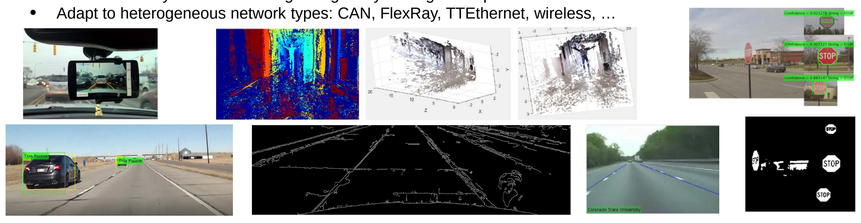
- **Photonic DRAM to CPU architectures**
  - **Faster transfers to support high density main memory**

Automotive Embedded Systems

- **Vehicles are controlled by distributed, real-time embedded systems**
  - Hundreds of embedded controllers/devices and millions of lines of code
  - Connected by multiple, diverse network protocols
  - **Challenge:**
    - Meet real-time computation and communication requirements
    - Prevent security breaches (tampering, snooping, …)
    - Support advanced driver-assistance systems (ADAS)



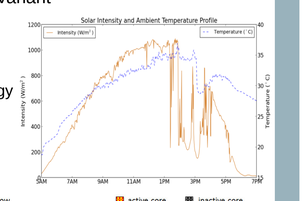
- **Jitter and security-aware automotive network design**
  - Dynamically adapt to unpredictable performance jitters (delays)
  - Prevent security breaches with lightweight key management protocols
  - Adapt to heterogeneous network types: CAN, FlexRay, TTEthernet, wireless, …

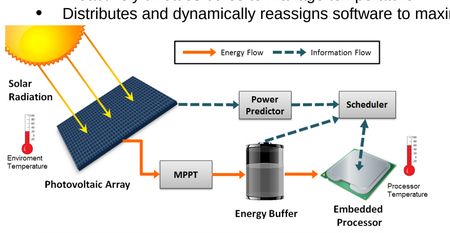


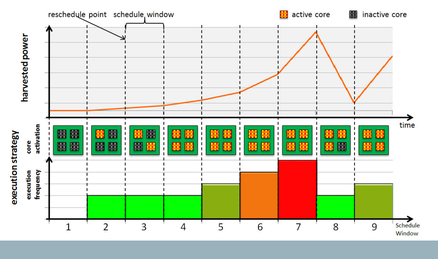
- **Advanced driver assistance system (ADAS) algorithms and prototyping**
  - Enable autonomous vehicles: design robust vehicle/pedestrian/traffic-light/sign/lane detection algorithms
  - Implement vision algorithms on low-power ADAS boards, smartphones, tablets
  - Utilize stereo vision cameras, and other data from: LIDAR, RADAR, vehicle-to-vehicle, vehicle-to-infrastructure, …

Energy Harvesting IoT Platforms

- **Solar energy harvesting** can power many IoT and embedded systems
  - How to **schedule software** applications on multicore platforms under variant and stringent energy harvesting conditions that often exist at runtime?
  - How to cope with **thermal spikes** and **faults** arising at runtime?
- **Run-time harvesting-aware scheduling framework**
  - Dynamically enables/disables cores, scales voltage/frequency to manage energy
  - Proactively throttles cores to manage temperature
  - Distributes and dynamically reassigns software to maximize core utilization



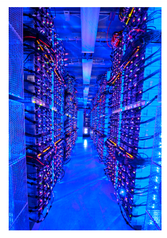




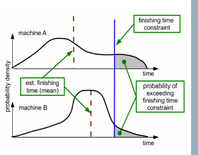
Thermal management Scheme	Average peak temperature	Number of throttlings	Overall task miss rate
throttling-only	79.60° C	94	35.92%
throttling + proactive	78.53° C	74	35.33%

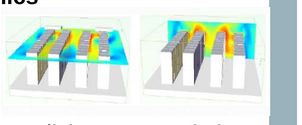
High Performance Computing

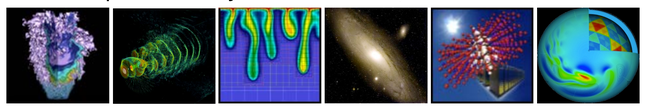
- **Many embedded systems use cloud computing**
- **Major challenge** in designing datacenters that support cloud computing as well as supercomputers that solve large scientific problems: **need for energy-efficient operation**
  - Energy costs today ~ $1M/year/petaflop
    - Cannot sustain such costs at exascale!
- **How can we reduce energy costs?**



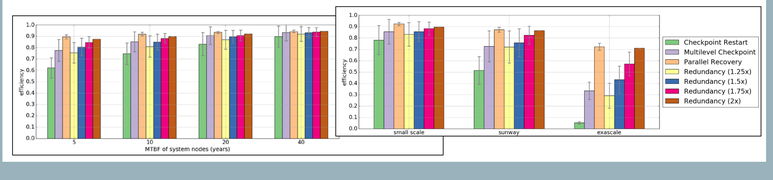
- **Energy Efficient and Stochastically Robust Resource Allocation**
  - **Workload and system uncertainty modeling**
    - Model uncertainty in execution time, network transfers, data access, and GPU offloading
    - Quantify task and machine heterogeneities in real world HPC systems
  - **Analysis of thermal and energy dynamics**
    - Real time thermal analysis
    - Adapting thermal setpoints
    - Characterize cooling energy & costs
  - **Smart resource allocation algorithms**
    - Workload, data, and storage allocation policies to co-optimize robustness, performance, and cooling/compute energy
      - Based on uncertainty models and thermal/energy analysis
    - Validation on diverse scientific applications and real world tera- and peta-scale systems at NCAR, DOE/ORNL, and CSU





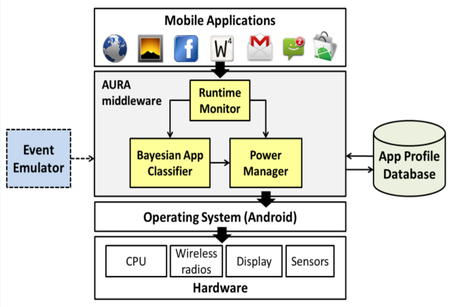


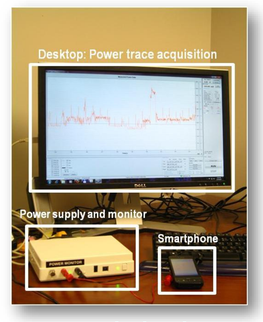
- **Another major challenge:** ensuring **fault-resilient operation**
  - Exascale HPC systems will experience a fault every few minutes!
  - How to quickly and effectively recover from frequent faults?
- **Reliability exploration/management for extreme-scale HPC**
  - **Analysis of checkpointing, redundancy based techniques**
  - **Co-design of resilience strategies with scheduling schemes**



Mobile Computing

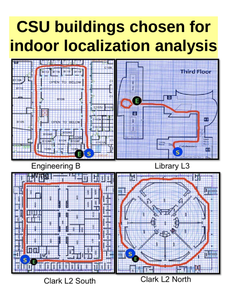
- **Energy demands** and **capabilities** of "smart" mobile devices are increasing rapidly with growing mobile app complexity
  - But **battery technology** is lagging behind and is expected to continue to be a limiting factor for future growth of mobile devices such as smartphones
  - How to intelligently **manage energy** and improve **battery lifetime** for mobile devices?



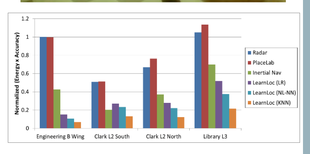


- **AURA middleware for CPU/backlight energy optimization**
  - Predicts idle periods (perceptual, cognitive, motor) during user-app interactions
  - Bayesian classification of mobile apps at runtime based on user-device interactions
  - Markov Decision Process (MDP) based algorithms to control:
    - dynamic voltage/freq. scaling (DVFS) for CPU energy saving during idle periods
    - backlight level (and energy) reduction based on theory of human change blindness
  - Power model based on real measurements of various Android OS-based smartphones
  - Avg. energy savings of **29%** vs. default Android scheme; **5x** over prior work; no QoS impact
- **Context-aware cloud offloading, wireless data transfers, and outdoor location sensing**
  - Reduce energy for data transmission and outdoor location sensing on mobile devices
  - Use software-based machine learning techniques to learn usage of data and location interfaces to determine optimal interface selection, ON/OFF schedule, configuration
    - Linear discriminant analysis, linear logistic regression, Non-linear logistic regression with neural networks, k-nearest neighbor, Support vector machines
  - Use similar techniques to determine when it is beneficial to compute in cloud vs. device
  - Up to **85%** energy savings vs. default Android scheme; **24%** savings over prior work

CSU buildings chosen for indoor localization analysis



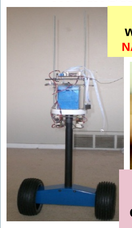
Localization accuracy variation with different fingerprinting capture frequencies



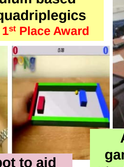
- **Energy-efficient and accurate indoor location sensing**
  - Indoor location sensing is difficult due to lack of GPS signals in indoor environments
  - Current techniques are energy-hungry, lack accuracy, and are very infrastructure dependent
  - Can use Wifi/UWB/cellular fingerprinting and inertial sensors to predict location indoors
  - Use machine learning techniques (LearnLoc) together with fingerprinting and inertial sensing (dead reckoning) to improve prediction accuracy and save energy
  - **Critical for search-and-rescue during emergency scenarios (e.g., cave-ins in mining)**

Embedded System Applications and Prototypes

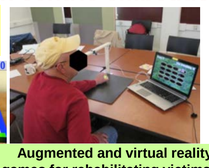
- **Medical and rehabilitation centric embedded systems**



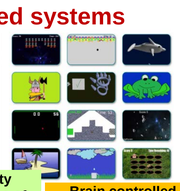
Inverted pendulum based wheelchair for quadriplegics
NASA Symposium 1st Place Award



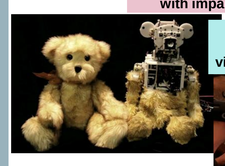
Mimicking robot to aid motor and speech development in children with impairments



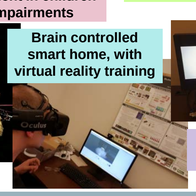
Augmented and virtual reality games for rehabilitating victims of stroke, cerebral palsy, and traumatic brain injury with motor disabilities

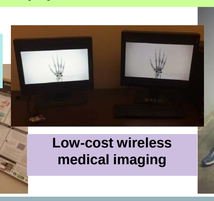


Brain controlled wheelchair

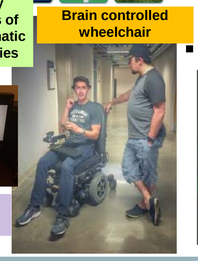


Brain controlled smart home, with virtual reality training





Low-cost wireless medical imaging



- **Mobile robotic embedded systems**

LARVA reconnaissance drone
2011 CSU E-days 1ST Place Award

Solar powered and networked terrain mapping robots

Hexapod sensing, fire monitoring and rescue assistance robot

LARRY autonomous retrieval robot
2012 CSU E-days 1ST Place Award

Remote wireless hacking drone

- **Other embedded applications**

Interactive guitar trainer

Laser engraving system

Smartphone controlled briefcase lock

RFID and Wi-Fi based home security system

Smartphone based automotive control platform

Smart mirror

Graduate Students

- **Current:** Sai Chittamuru, Ishan Thakkar, Daniel Dauwe, Yaswanth Raparti, Vipin Kukkala, Saideep Tiku, Shoumik Maiti, Greg Kittilson, Ninad Hogade, Yahav Biran, Chris Langlois, Varun Kilenje, Swapnil Bhosale, Ayush Kumar, Jordan Tunnell, Zemin Tao, Rohit Kudre, Rohan Jhaveri
- **Alumni:** Shirish Bahirat, Yong Zou, Yi Xiang, Nishit Kapadia, Mark Oxley, Brad Donohoo, Pramit Rajakrishna, Miguel Salas, Tejasi Pimpalkhute, Viney Ugave, Eric Jonardi, Yuhang Li, Srinivas Desai, Haneet Mahajan, Aditya Khune, Onkar Gulvani, Jiabao Jin, Manoj Kumar, Sai Kiran, Nanda Kumar, Taylo Santiago, Surya Vamsi Vemparala, Jingjie Zhu