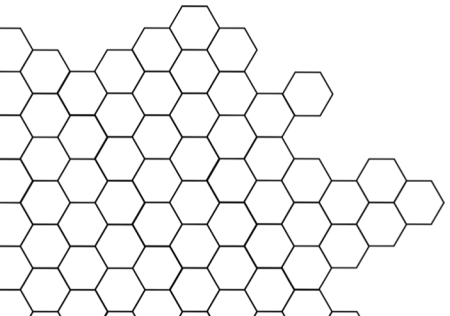
ILLINOIS INSTITUTE OF TECHNOLOGY

Overview of the HExSA Lab @ IIT

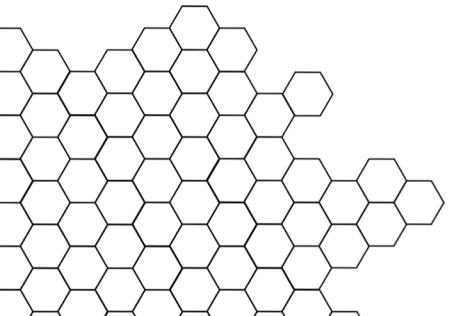
Laboratory for High-performance Experimental Systems and Architecture

PI: Kyle Hale

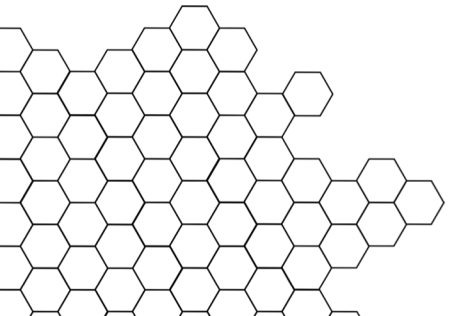


Three Primary Themes

- **High-performance Operating Systems,** runtime systems, and virtual machines
- **Novel languages and runtimes** for parallel and experimental systems
- **Experimental computer architectures**



Current thrusts



ILLINOIS INSTITUTE OF TECHNOLOGY

High-performance Operating Systems and Virtual Machines

- **Nautilus** and Hybrid Runtimes (**with Prescience Lab @ Northwestern**)
- Compiler + Kernel fusion [**The Interweaving Project**] (**with CS groups @ Northwestern**)
- Hybrid Runtime for Compiled Dataflows [**HCDF**] (**with DBGroup @IIT**)
- Address Space Dynamics
- **High-performance Virtualization** [Palacios VMM[3] and Pisces Cokernels[4]] (**with Prescience Lab @ Northwestern; Prognostic Lab @ Pitt**)
- High-performance networking
- Accelerated Asynchronous Software Events [**Nemo**]
- Computational Sprinting (with U. Nevada, Reno and OSU)

Nautilus and HRTs

ILLINOIS INSTITUTE OF TECHNOLOGY

- High-performance ***Unikernel for HPC, parallel computing***[1]
- ***Hybrid Runtime (HRT)***[2] = parallel runtime system + kernel mashup
- Lightweight, fast, single-address space Operating System
- ***Designed to make parallel runtimes efficient and well-matched to the hardware***
- Sponsored by NSF, DOE, and Sandia National Labs
- Collaboration with Prescience Lab[3] at Northwestern

Northwestern University

HOBBES
xstack.sandia.gov/hobbes

Sandia National Laboratories

[1]http://presciencelab.org
[2]http://nautilus.halek.co
[3]http://users.eecs.northwestern.edu/~kch479/docs/nautilus.pdf

The Interweaving Project[1]

- Unikernels provide a new opportunity for *combining kernel, user, and runtime code*
- ***Interweave*** them into one binary
- Compiler generates OS code, driver code
- ***Compiler/Runtime/OS/Architecture Co-Design***
- Collaboration with Prescience Lab, PARAG@N Lab, and Campanoni Lab @ Northwestern
- NSF sponsored, $1M, 4 PIs

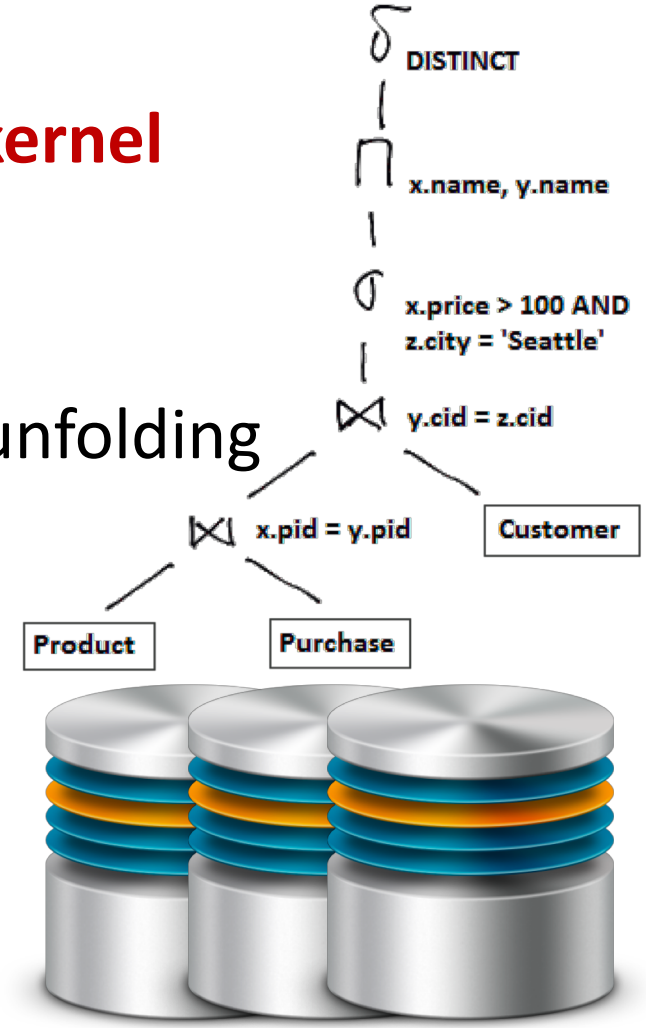
[1]http://interweaving.org

Northwestern University

ILLINOIS INSTITUTE OF TECHNOLOGY

Hybrid Runtime for Compiled Dataflows (HCDF)

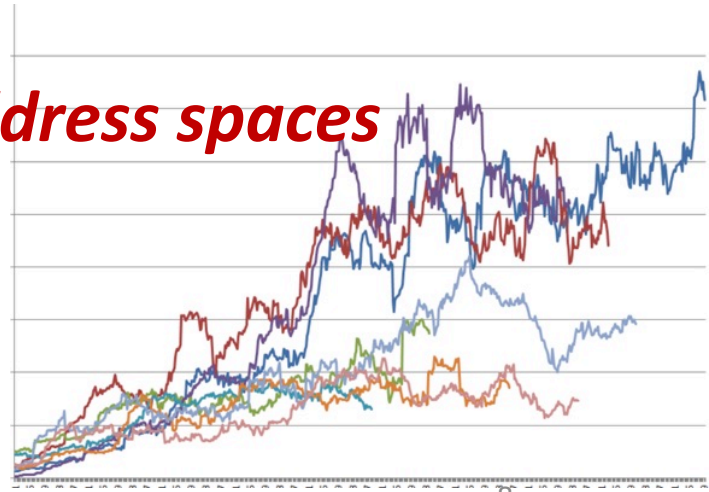
- **Co-Design database engine and operating system kernel**
- Compiled queries placed into tasks, scheduled onto specialized hybrid runtime in an OS kernel
- **Runtime extracts parallelism and performance** by unfolding query task graph and tailored hardware access
- Collaboration with DB Group @ IIT



DBGroup

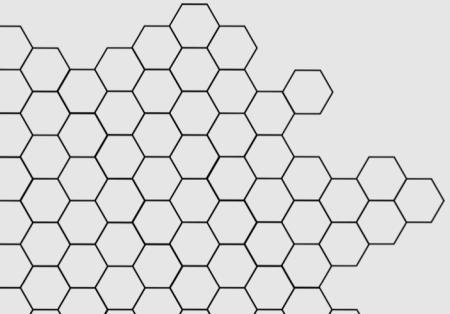
Address Space Dynamics

- Ubiquitous virtualization is putting pressure on address translation hardware and software
- New chip designs also pressing the issue (5-level PTs in next-gen Intel chips)
- We're looking at ***new address translation mechanisms*** (Interweaving Project)
- These may require understanding ***the structure of address spaces over time***
- ***Can we discover this dynamic structure?***



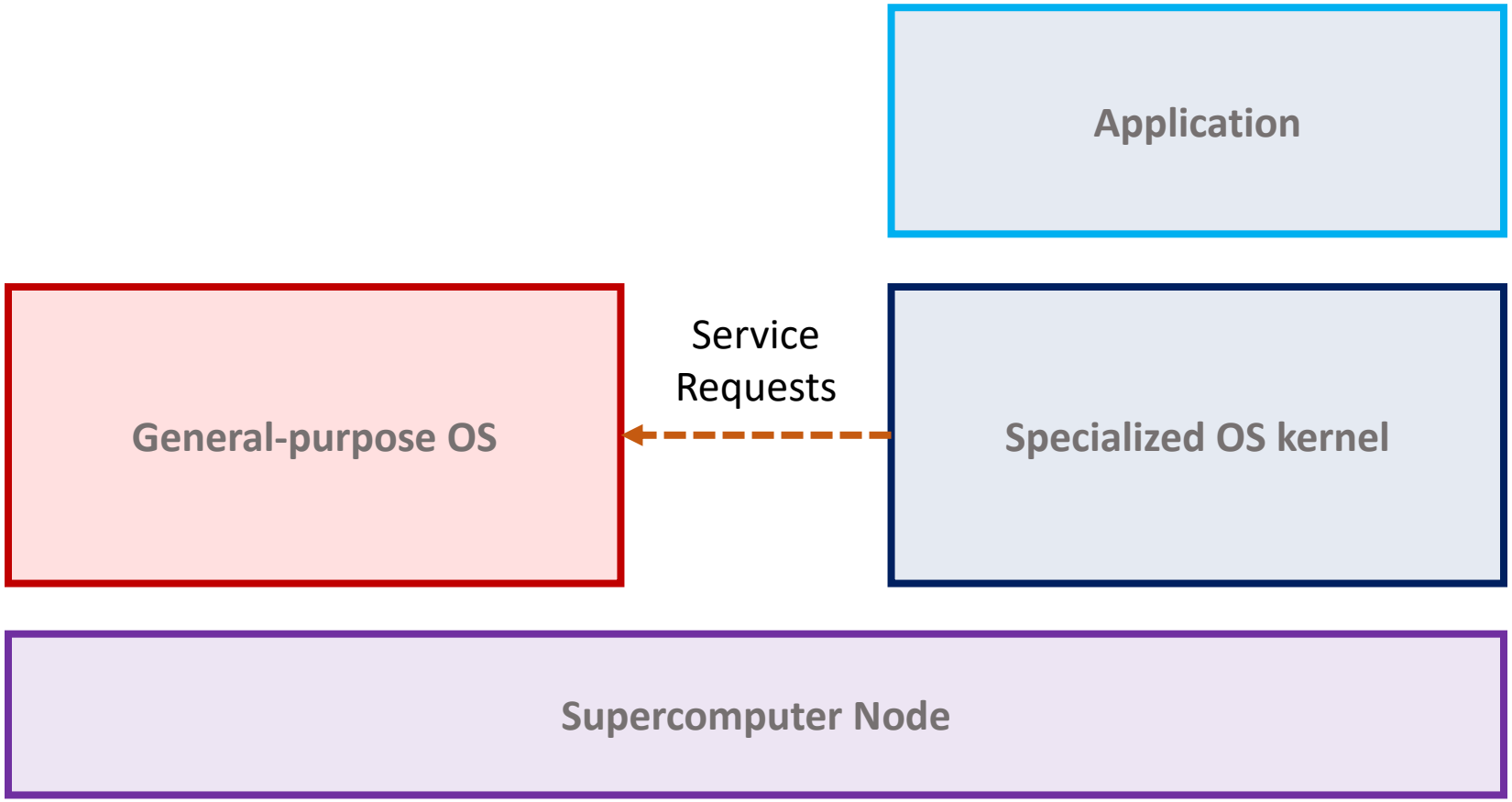
Multi-kernel Systems for Supercomputing

- **Hybrid Virtual Machines[1]** [multi-kernel VMs]
- **Multiverse:** run legacy apps. on a multi-kernel VM
- Modeling system call delegation [**Amdahl's Law for multikernels**]
- **High-performance Virtualization** [Palacios VMM and Pisces Cokernels]
- Coordinated kernels as containers [**SOSR Project**]



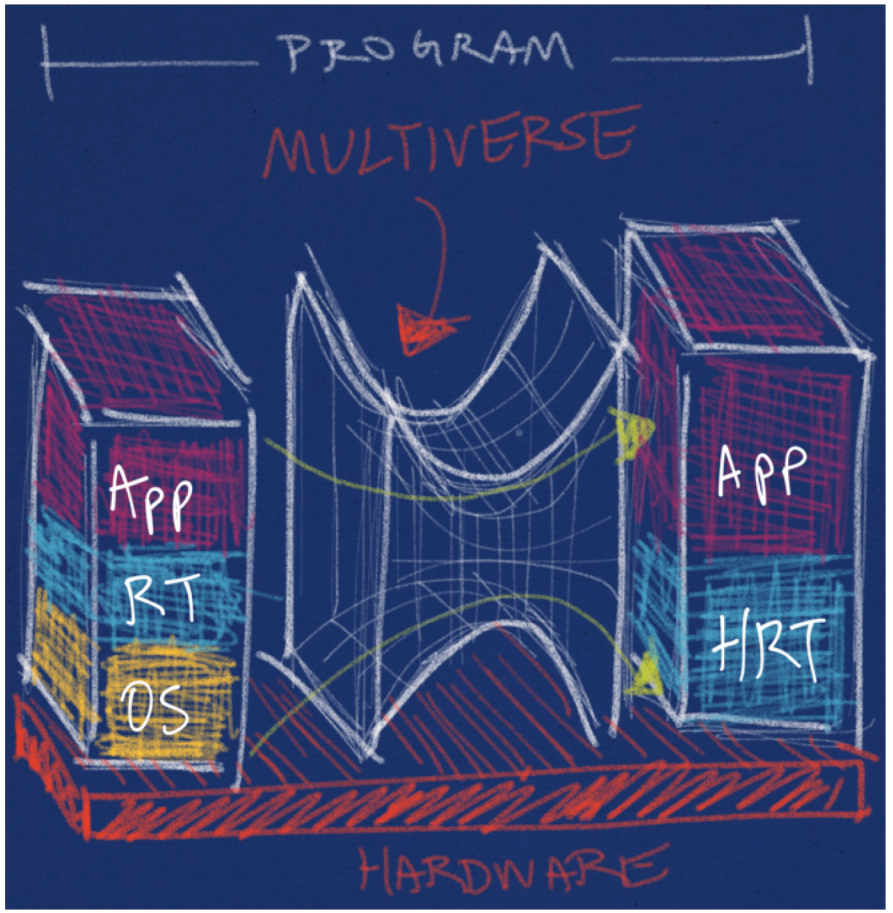
[1]http://pdinda.org/Papers/vee16.pdf

The Multikernel Approach



Multiverse[1]

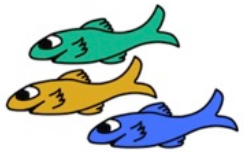
- Typically must *port* your parallel program to run in Multikernel environment
- **We automatically port legacy apps to run in this mode**
- Uses a ***virtualized* multikernel** approach
- Working example with the Racket[2] runtime system



[1]http://cs.iit.edu/~khale/docs/icac17-multiverse.pdf
[2]https://racket-lang.org

Coordinated SOS/Rs for the Cloud

- Specialized Operating Systems and Runtimes (SOS/Rs) (e.g. Unikernels) are difficult to use!
- Leverage programming model and interface of ***containers*** to ease this problem => ***Containerized Operating Systems***
- Treat a collection of SOS/Rs within a single machine as a distributed system (requires coordination)
- Collaboration with Prognostic Lab @ Pitt
- NSF-sponsored, $500K (2 PIs)



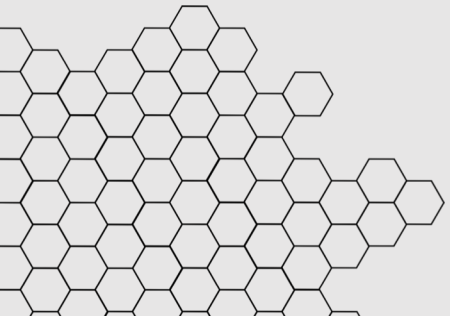
Pisces
Isolated Lightweight Co-kernels

University of Pittsburgh

Novel Languages and Runtimes for Parallel and Experimental Systems

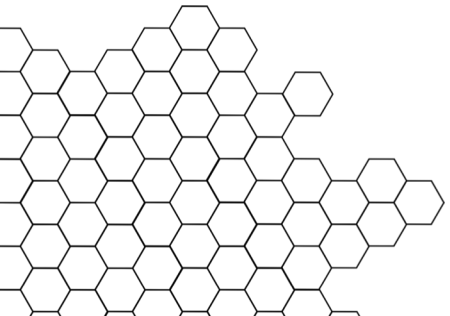
ILLINOIS INSTITUTE OF TECHNOLOGY

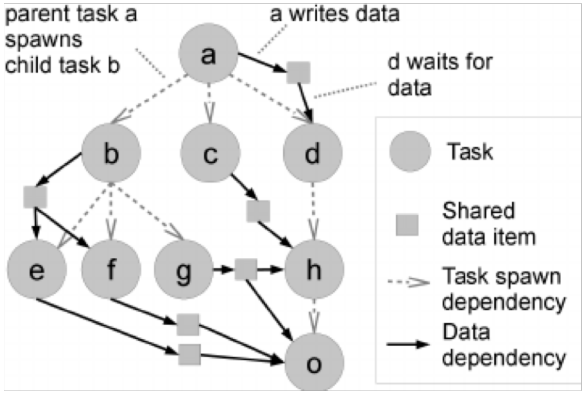
- Exploration of **Julia for large-scale**, parallel computing
- **XTask** – A runtime system for extrem-scale, fine-grained, many-task computing (**with DataSys Lab @IIT**)
- **New systems languages**
- **New virtual machine architectures** for dataflow-oriented programming models (virtual, spatial computing)



XTask

- **Future supercomputers will have millions and millions of short, *fine-grained* tasks** (think user/green threads)
- Current tasking runtimes assume long-running, computation heavy tasks
- How do we **build efficient, low-overhead runtimes to support this**?
- Collaboration with DataSys Lab @ IIT and Prescience Lab @ Northwestern



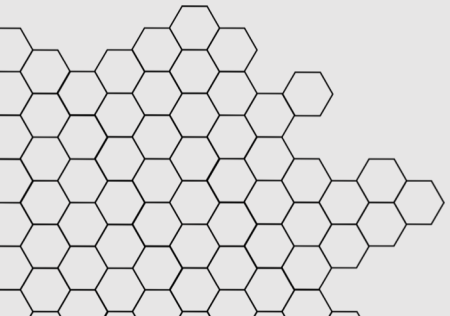


DataSys
Data-Intensive Distributed Systems Laboratory

Plab

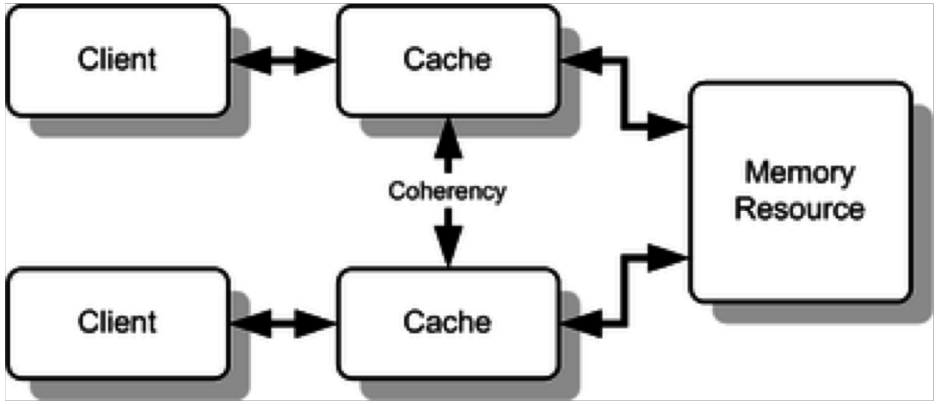
Experimental Computer Architectures

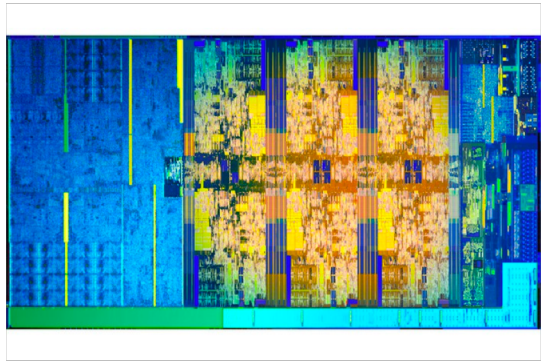
- **State-associative prefetching**: using neuromorphic chips to prefetch data between levels of deep memory hierarchies
- **DSAs for Hearing** Assistance [**with collab. at Interactive Audio Lab @ Northwestern**]
- **Incoherent Multicore Architectures** (**with CS @ Northwestern**)



Incoherent Multicore Architectures

- The cost of cache coherence (keeping local caches consistent in multi-cores) goes up with scale
- **Certain software doesn't need it, but pays for its effects**
- ***Can we get rid of it?*** What would software-managed coherence look like?

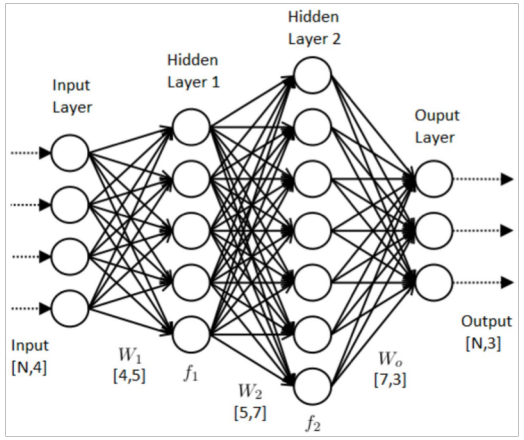


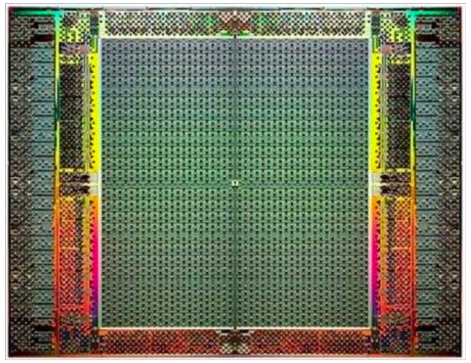


PARAGON
PARALLEL ARCHITECTURE GROUP @NORTHWESTERN

Domain-Specific Architectures for Hearing Assistance

- "**Cocktail problem**": Identify speaker in a crowded (loud) room
- **Brain is very good at this**
- **Hearing aids *are not*** (they typically apply some pretty simple signal processing)
- We're looking to design a **new chip architecture for hearing aids based on audio source separation** (a machine learning-based technique)





"Out there" stuff

- **"Parsec-scale" parallel computing**
- Exploring the **kinematics of execution contexts** ("can you use a Lagrangian to describe processes as a dynamical system?")
- **Decentralized hash algorithm evaluation** and verification "hashes for the masses"

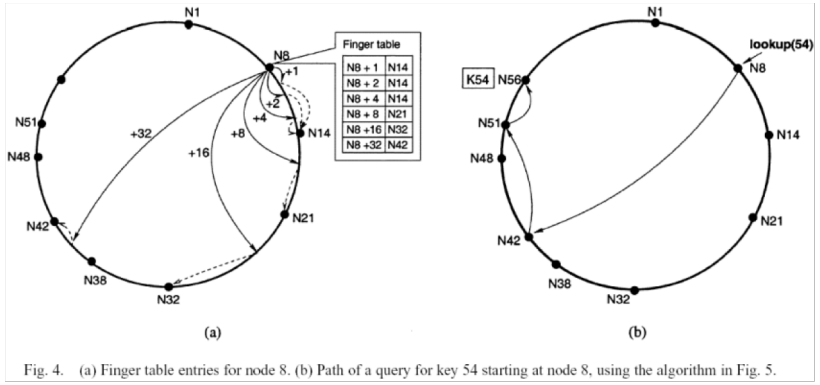
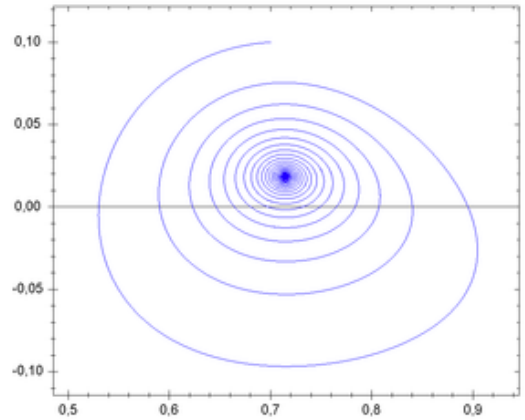
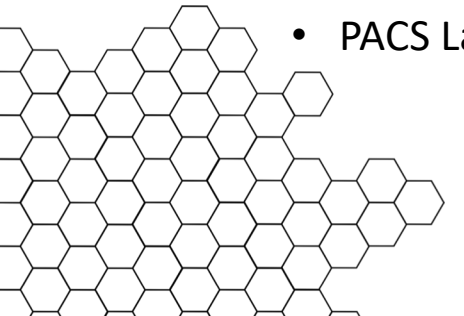


Fig. 4. (a) Finger table entries for node 8. (b) Path of a query for key 54 starting at node 8, using the algorithm in Fig. 5.



Collaborators

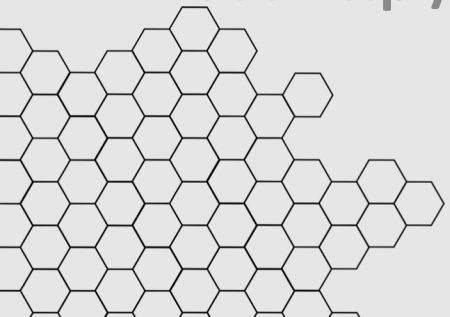
- **IIT**
  - Scalable Systems Laboratory (Xian-He Sun)
  - DB Group (Boris Glavic)
  - DataSys Lab (Ioan Raicu)
- **Northwestern University**
  - Prescience Lab (Peter Dinda)
  - PARAG@N Lab (Nikos Hardavellas)
  - Campanoni Lab (Simone Campanoni)
- **University of Pittsburgh**
  - Prognostic Lab (Jack Lange)
- **Ohio State University**
  - ReRout Lab (Christopher Stewart)
  - PACS Lab (Xiaorui Wang)
- **University of Nevada @ Reno**
  - IDS Lab (Feng Yan)
- **University of Chicago**
  - Kyle Chard
  - Justin Wozniak
- **Sandia National Laboratories**
  - Kevin Pedretti
- **Pacific Northwest National Laboratories**
  - High Performance Computing Group (Roberto Gioiosa)



We're hiring!

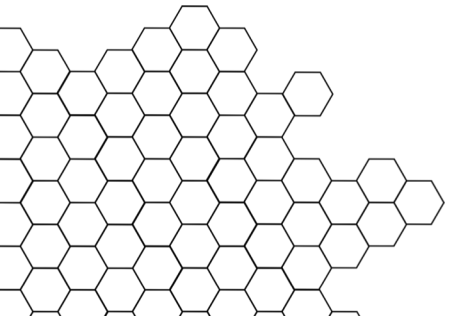
Funded opportunities available (both PhDs and undergrads!)

See **http://cs.iit.edu/~khale/student_apps.html**



Relevant Courses

- **CS 450**: Operating Systems
- **CS 562**: Virtual Machines (was formerly CS 595 F17, F18)
- **CS 595-03**: OS and Runtime Design for Supercomputing (Research Seminar)
- **CS 551**: Operating System Design and Implementation (grad OS, I'm not teaching this yet)



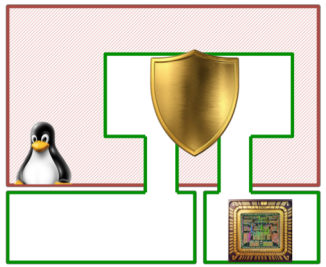
Completed Projects

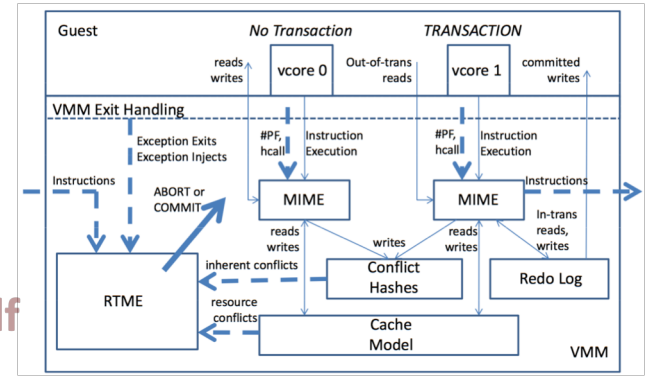
- Philix Xeon Phi OS Toolkit[1]
- Palacios VMM[2]
- Guest Examination and Revision Services (GEARS[3])
- Guarded Modules[4]
- Virtualized Hardware Transactional Memory[5]

philix
boot your third-party OS on the Intel Xeon Phi

V3VEE

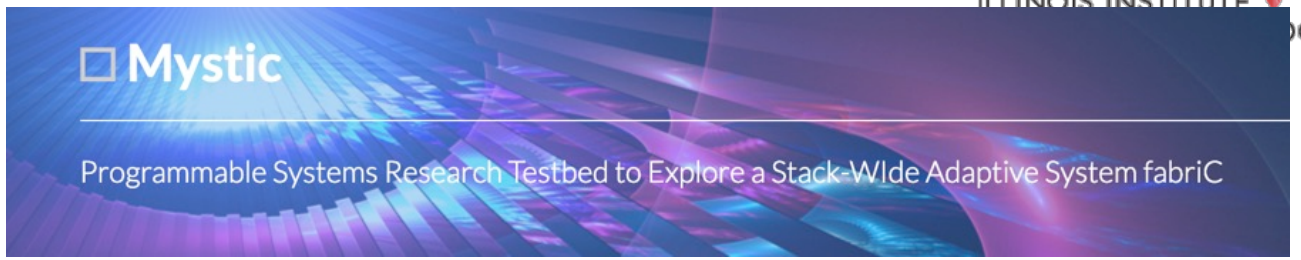
Palacios
An OS Independent Embeddable VMM





[1]http://philix.halek.co
[2]http://v3vee.org/palacios
[3]http://users.eecs.northwestern.edu/~kch479/docs/gears.pdf
[4]http://users.eecs.northwestern.edu/~kch479/docs/gm.pdf
[5]http://users.eecs.northwestern.edu/~msw978/resources/palacios-htm.pdf

Cool hardware



- **HExSA Rack**
  - Newest Skylake and AMD Epyc machines (may-core)
  - Designed for booting OSes
- **Supercomputer Access**
  - Stampede2 Supercomputer @ TACC
  - Comet Cluster @ SDSC
  - Jetstream Supercomputer @ IU
  - Chameleon Cloud

- **MYSTIC Cluster**
  - 8 Dual Arria 10 FPGA systems
  - 8 Mellanox Bluefield SoC systems
  - Newest ARM servers
  - IBM POWER9
  - Xeon Scalable Processor systems
  - 16 NVIDIA V100 GPUs
  - 100Gb internal network (Infiniband and 10GbE)