HAJIM
SCHOOL OF ENGINEERING
& APPLIED SCIENCES
UNIVERSITY of ROCHESTER

DEPARTMENT OF
COMPUTER SCIENCE

Timescale Stream Statistics for Hierarchical Management

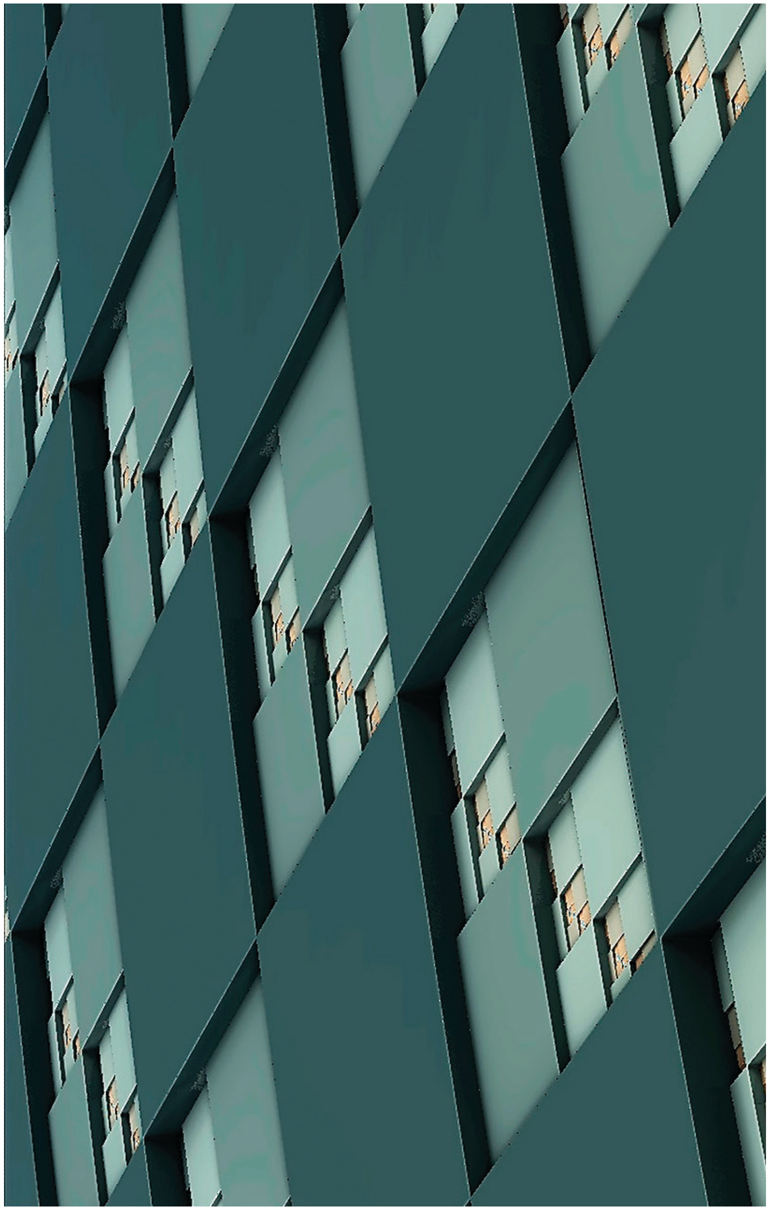
Chen Ding
University of Rochester

March 23
STREAM 2016
Tysons, VA

Implications of the datacenter's shifting center.

BY MIHIR NANAVATI, MALTE SCHWARZKOPF, JAKE WIRES, AND ANDREW WARFIELD

Non-Volatile Storage



"The arrival of high-speed, non-volatile storage … is likely the most significant architectural change that datacenter and software designers will face in the foreseeable future."

Hierarchical Cache Memory

- Science
  - nothing travels faster than light
    - the faster the access, the smaller the data capacity
- Engineering
  - speed, size and cost
    - no single technology can satisfy all demands
  - e.g. SCM mentioned in the CACM article
- Programmability
  - automatic, transparent, modular, efficient, portable
  - efficient sharing of fast/local memory
- Uses
  - CPU/GPU caches, virtual memory
  - software cache, e.g. Memcached, Redis

GPU	G80	GT200	Fermi
Transistors	681 million	1.4 billion	3.0 billion
CUDA Cores	128	240	512
Double Precision Floating Point Capability	None	30 FMA ops / clock	256 FMA ops /clock
Single Precision Floating Point Capability	128 MAD ops/clock	240 MAD ops / clock	512 FMA ops /clock
Special Function Units (SFUs) / SM	2	2	4
Warp schedulers (per SM)	1	1	2
Shared Memory (per SM)	16 KB	16 KB	Configurable 48 KB or 16 KB
L1 Cache (per SM)	None	None	Configurable 16 KB or 48 KB
L2 Cache	None	None	768 KB
ECC Memory Support	No	No	Yes
Concurrent Kernels	No	No	Up to 16
Load/Store Address Width	32-bit	32-bit	64-bit

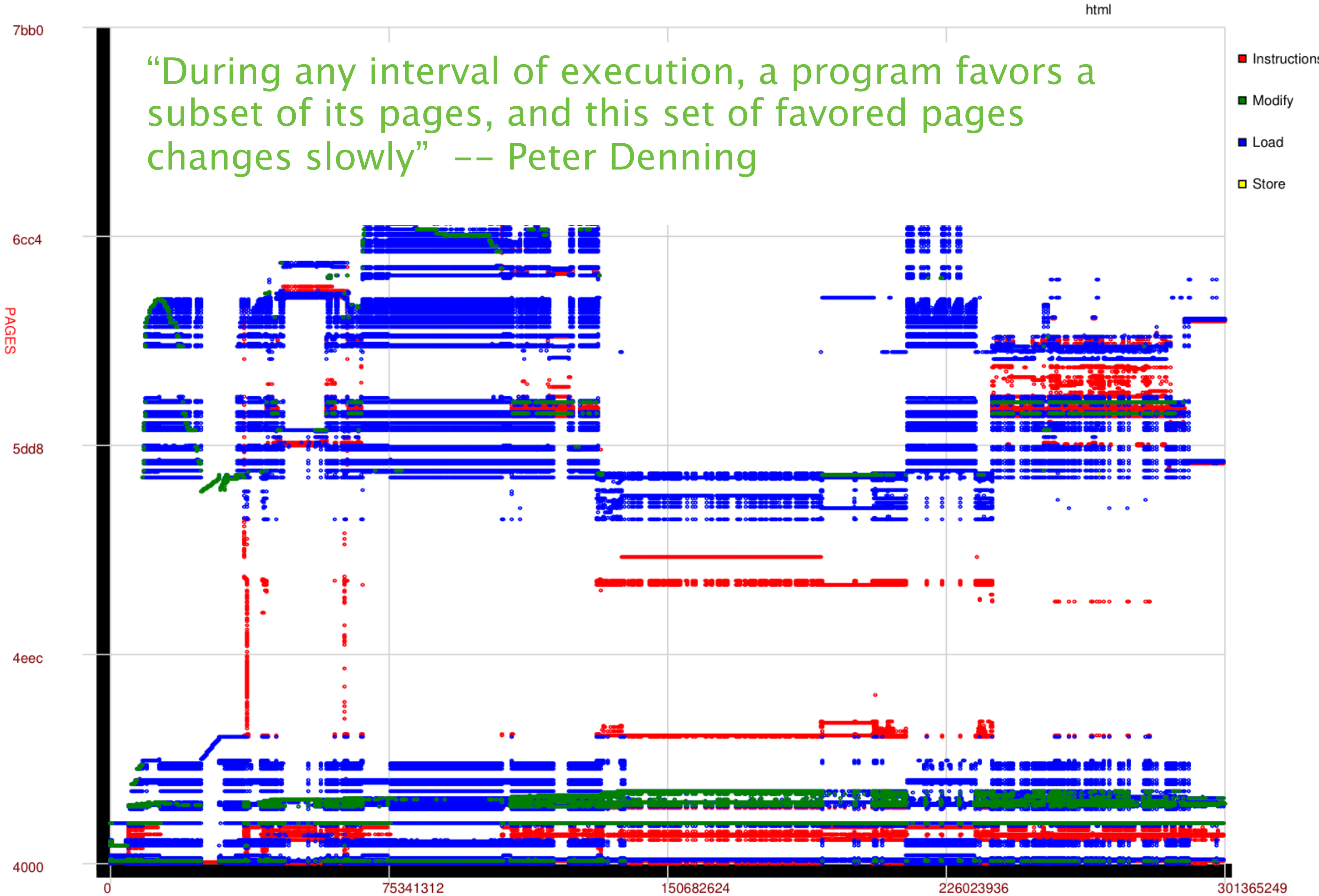
Whitepaper

NVIDIA's Next Generation CUDA™ Compute Architecture:

Fermi™

What is Locality?

"During any interval of execution, a program favors a subset of its pages, and this set of favored pages changes slowly" -- Peter Denning



- locality analysis is a streaming problem
- too many data points, unusable for optimization

Locality Theory

- Since 1960s
  - working-set theory [Denning 1968]
  - stack simulation [Mattson et al. 1970]
- Since 1999
  - reuse distance (i.e. LRU stack distance)
  - 5 dimensions of locality [TOPLAS'09]

  "The authors were supported by the National Science Foundation (CAREER Award CCR-0238176 and two grants CNS-0720796 and CNS-0509270), the Department of Energy (Young Investigator Award DE-FG02-02ER25525), IBM CAS Faculty Fellowship, and a gift from Microsoft Research. "

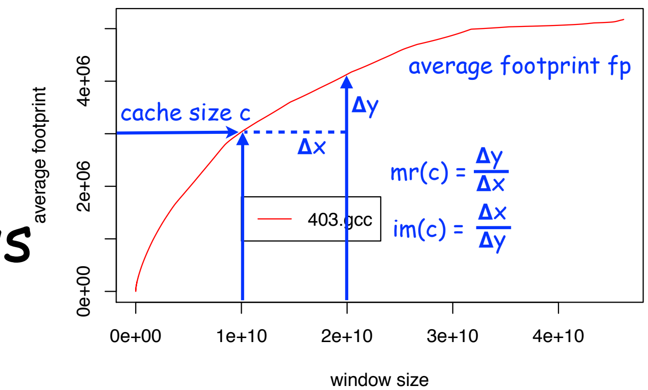
  - HPCToolkit by Mellor-Crummey et al. at Rice [CCPE'10]
  - not composable, unable to derive shared-cache performance
- Since 2008
  - footprint — timescale statistics

Timescale Stream Statistics

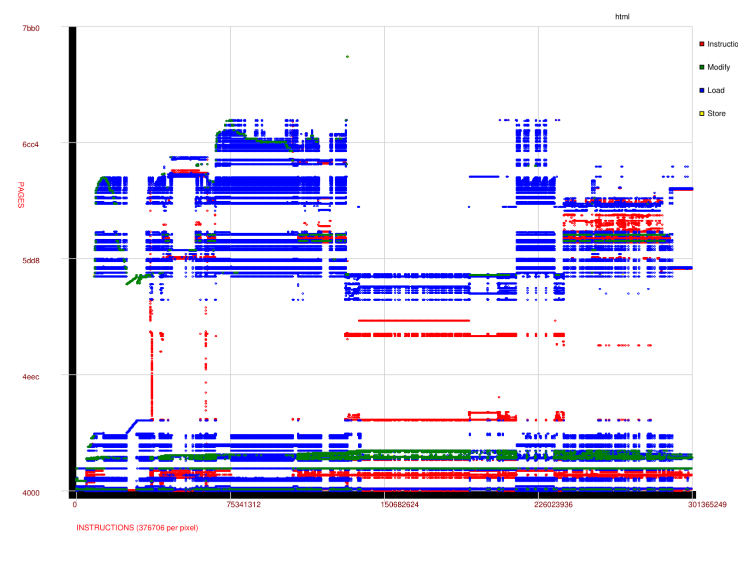
- A stream
  - "a possibly unbounded sequence of events" [Stream workshop 2015]
  - a time window or interval
  - a timescale x is a length of time
  - f(x) is the average behavior of all windows of length x
    - a function for all non-negative x
- Peak temperature variation pv(x)
  - each window has a peak variation
  - pv(x) is the average of all windows of length x
    - e.g. a week time or a month time
  - avoid data bias
    - e.g. if we were to measure just calendar weeks/months

Timescale Locality

- Footprint fp(x)
  - working-set size (WSS): the amount data accessed in a window
  - fp(x): average WSS of all length x windows



- Theoretical properties (selected)
  - composable
  - miss ratio is the increase of footprint
  - concavity [ASPLOS'13]
    - (computed) miss ratio is monotone
  - linear time measurement [PACT'11]
    - real-time sampling [CCGrid'15]
- A function is worth a thousand pictures



Theory is for Optimization

- Key-value store Memcached [USENIX'15]
  - DRAM as cache for database
  - optimization vs. heuristics by Facebook and Twitter
    - faster steadystate/convergence on a Facebook test set
  - monotonicity: no Belady anomaly
- Concurrent memory allocation [see white paper]
  - optimization vs. Google's tcmalloc
    - 26% higher throughput 64-thread MongoDB
  - consistency: intermediate steps order insensitive
- Storage cache [Wires/Warfield et al. OSDI'14]
  - independent validation of footprint theory
- Other theories
  - optimal data placement [PLDI'04, POPL'06, POPL'16]
  - optimal collaborative caching [LCPC'08, ISMM'11/12/13]

Summary: Locality Theory/Optimization

- Locality theory
  - partly a streaming problem/solution
  - equivalent* definitions of locality
    - reuse distance, footprint, working set, miss ratio curve
- Possible uses in a streaming system
  - Nathan's IPPD
    - memory resource steering
  - timescale statistics
    - user decision support
- A conjecture
  - memory: hierarchical and shared
  - timescale stream statistics: optimal sharing of a hierarchy