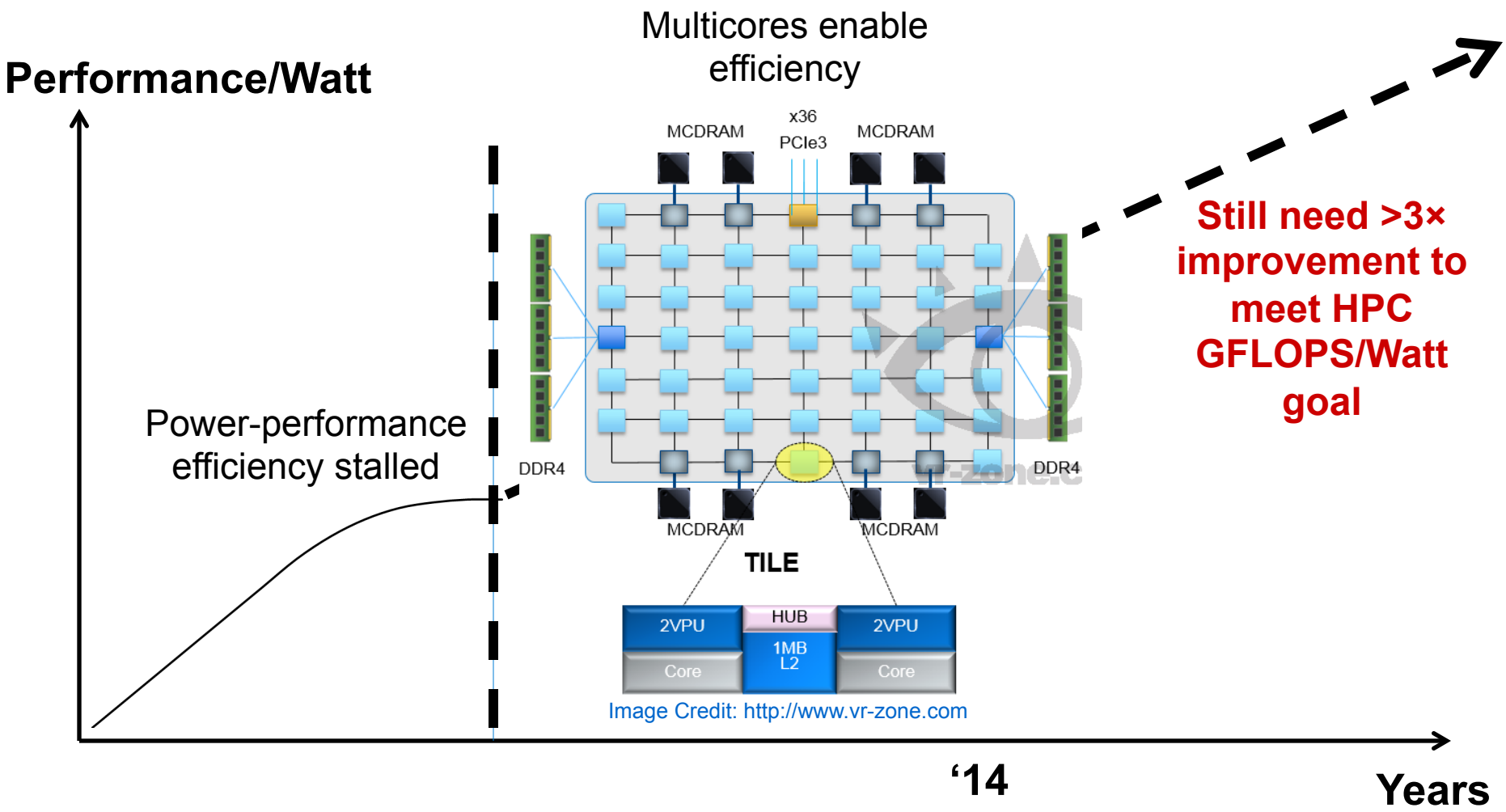
UCONN COMPUTER ARCHITECTURE GROUP

Rethinking Last-Level Cache Management for Multicores Operating at Near-Threshold

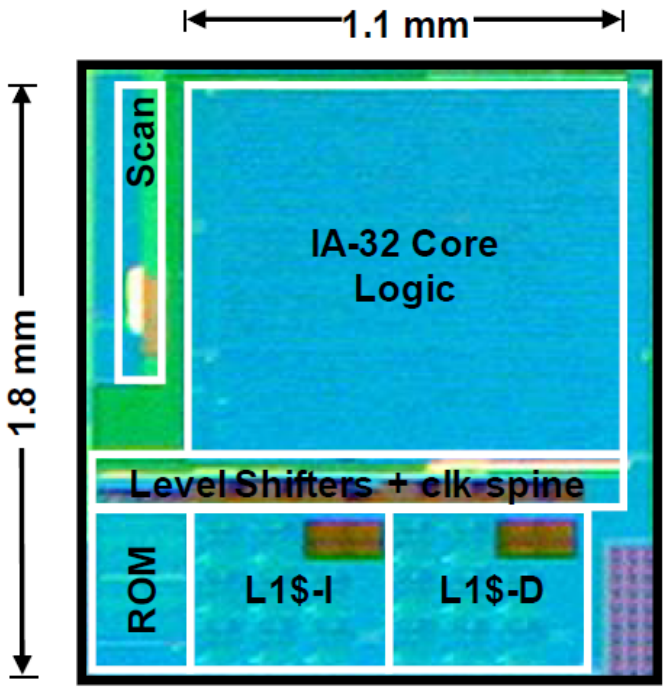
Farrukh Hijaz, Omer Khan
University of Connecticut

Power Efficiency



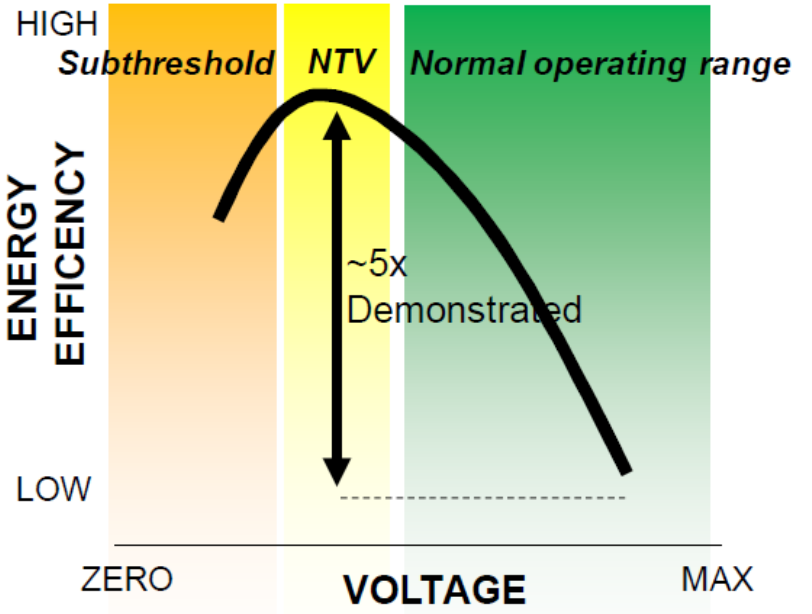
The Value of Operating at NTV

Near Threshold Voltage operation potentially enables 5-10× power-performance efficiency



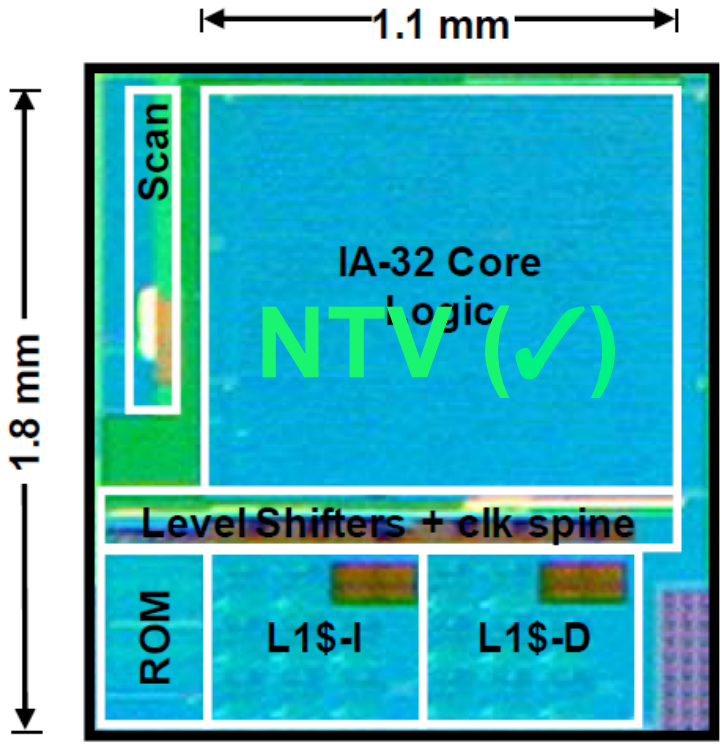
Pentium®, 32 nm CMOS

[Intel: DAC'12]



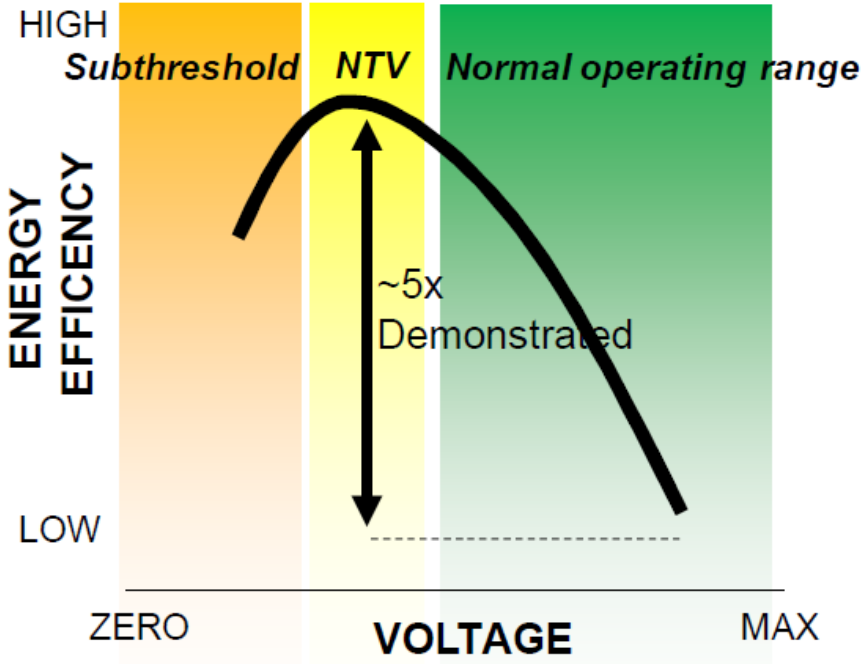
Ultra-low Power	Energy Efficient	High Performance
280 mV	0.45 V	1.2 V
3 MHz	60 MHz	915 MHz
2 mW	10 mW	737 mW
1500 Mips/W	5830 Mips/W	1240 Mips/W

NTV Operation? Logic (✓)



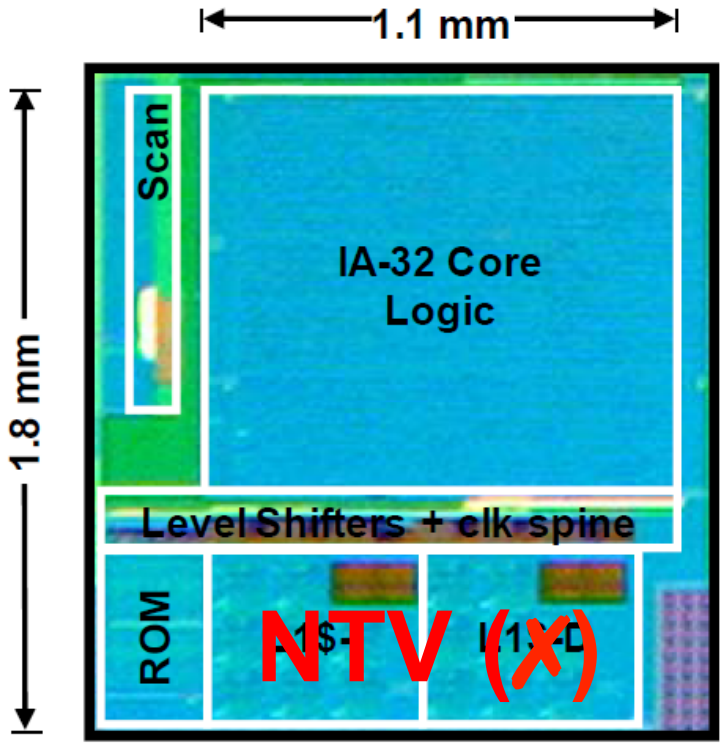
Pentium®, 32 nm CMOS

[Intel:DAC'12]



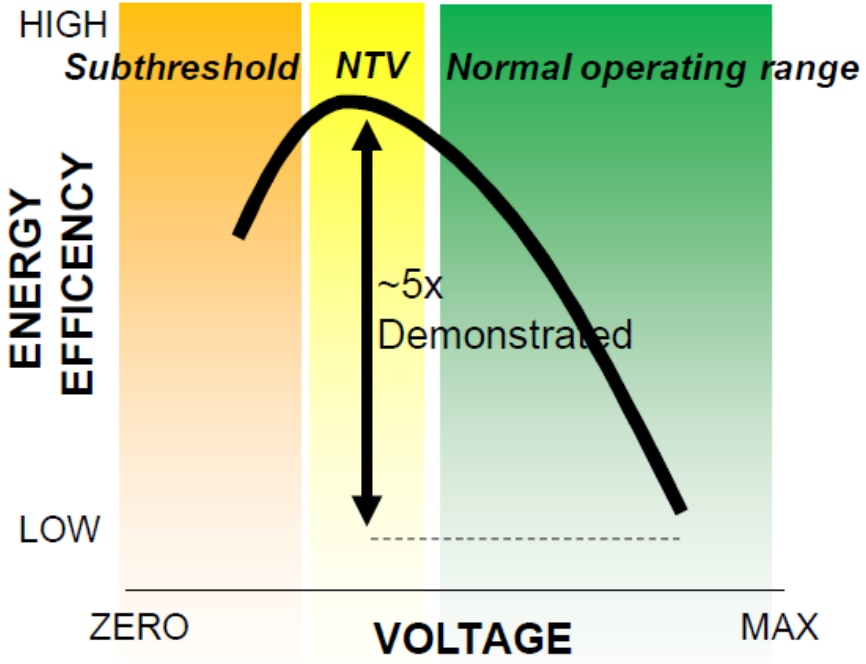
Ultra-low Power	Energy Efficient	High Performance
280 mV	0.45 V	1.2 V
3 MHz	60 MHz	915 MHz
2 mW	10 mW	737 mW
1500 Mips/W	5830 Mips/W	1240 Mips/W

NTV Operation? Cache (✗)



Pentium®, 32 nm CMOS

[Intel:DAC'12]



Ultra-low Power	Energy Efficient	High Performance
280 mV	0.45 V	1.2 V
3 MHz	60 MHz	915 MHz
2 mW	10 mW	737 mW
1500 Mips/W	5830 Mips/W	1240 Mips/W

SRAM bit-cells susceptible to errors at NTV

NTV Approaches for On-chip Memory

- High voltage, High frequency
  - High performance
  - Low energy efficiency
  - No faults
- Low voltage, Low frequency
  - Low performance
  - Highest energy efficiency
  - No faults
- Low voltage, High frequency **Our Approach!**
  - High performance
  - High energy efficiency
  - **Permanent faults**

NTV Approaches for Permanent Faults

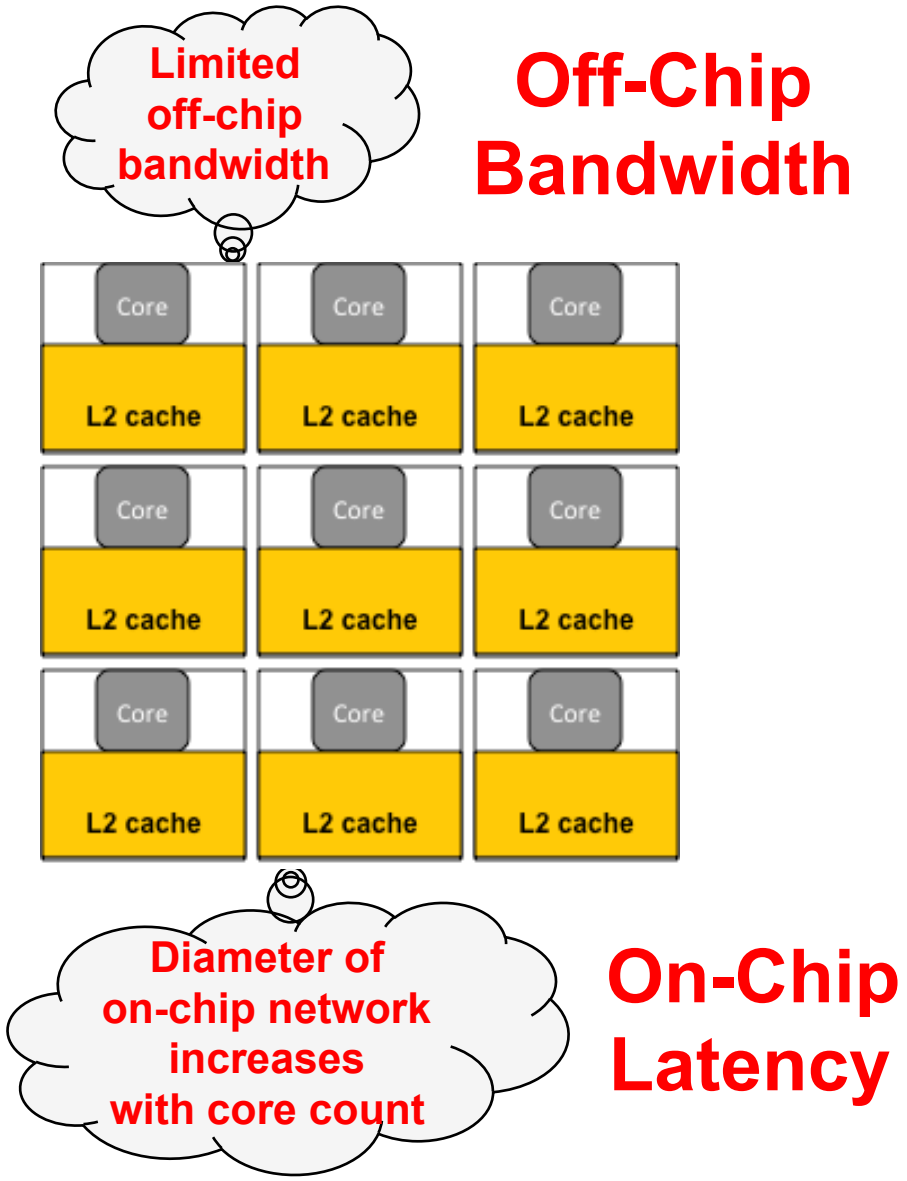
- Circuit level (8T, 10T SRAM bit-cell)
  - High area overhead
  - Higher leakage current
- ECC based (SECDED, MS-ECC)
  - Constant latency overhead
- Disabling based (e.g., cache line disabling)
  - Lower available capacity

Our Approach!

- **Hybrid of ECC and Disabling (e.g., VS-ECC)**
  - **Trades off available capacity and latency overhead**

The NTV Challenge in Multicores

- Future multicores will have 100s of cores
- LLC management is key to optimizing performance and energy
- Last-level cache (LLC) data locality and off-chip miss rates 1st order constraints and often show opposing trends
- **Lower available LLC capacity at NTV presents new challenges**



Static-NUCA
(LLC Data Placement)

- Statically address interleaves data across all physically distributed LLC slices
- No replication of data in the LLC slices
  - High cache utilization since all data evenly distributed
- Data resides in a remote LLC slice with high probability
  - High remote LLC slice access rate results in higher on-chip network traffic and high average LLC access latency/energy

Reactive-NUCA
(LLC Data Placement, Limited Replication)

- Classifies data as private or shared on page granularity using the existing virtual memory system
  - Maps *private* pages to requesting core's local LLC slice
  - Maps *shared* pages across the chip based on static address interleaving (similar to Static-NUCA)
- Replication of data not allowed
- Instructions replicated in LLC slice per cluster of 4, using rotational interleaving
- Low LLC access latency/energy for correctly classified private data and instructions
- No locality optimizations for shared data

Victim Replication
(LLC Data Placement and Replication)

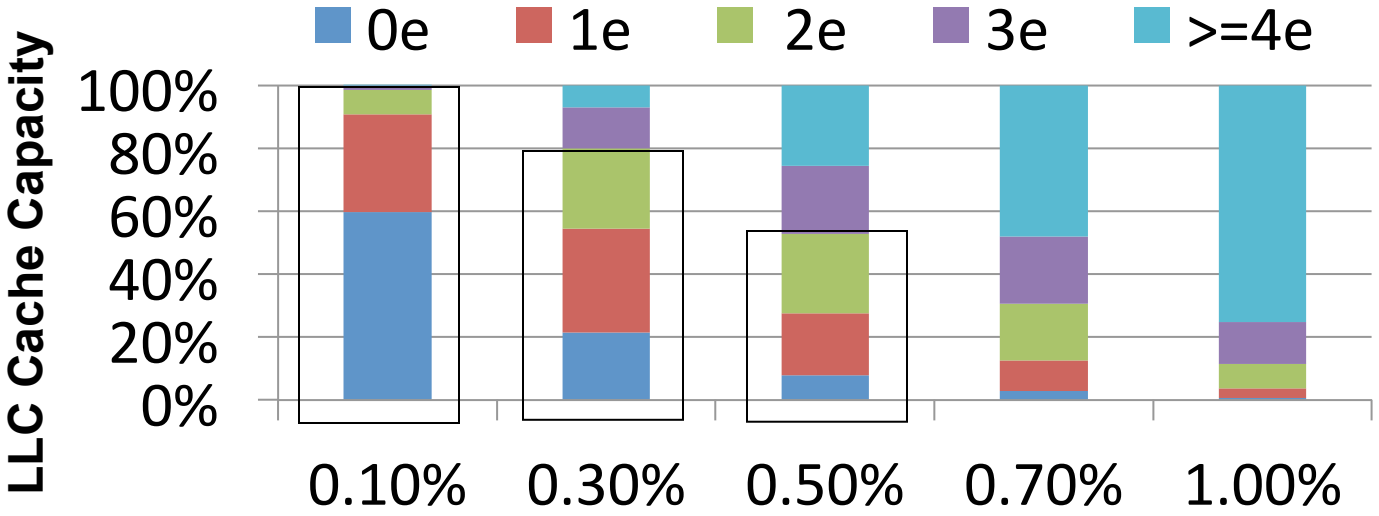
- Starts with S-NUCA and uses the local LLC slice of a core as a victim cache for the cache lines evicted from its L1 cache
- Inserts replica only if there exists:
  - an invalid cache line,
  - a home cache line with zero sharers, or
  - another replica
- Improves locality and reduces on-chip traffic
- Replication strategy causes LLC pollution, resulting in higher evictions of home cache lines with zero sharers and other replicas

Evaluation Methodology

- Evaluation using Graphite multicore simulator for **64 cores**
  - McPAT/CACTI cache energy models and DSENT network energy models at **11 nm**
- Evaluated **21 benchmarks** from the SPLASH-2 (11), PARSEC (8), Parallel MI-bench (1) and UHPC (1) suites
- LLC managements schemes compared:
  - Static-NUCA (S-NUCA)
  - Reactive-NUCA (R-NUCA)
  - Victim Replication (VR)

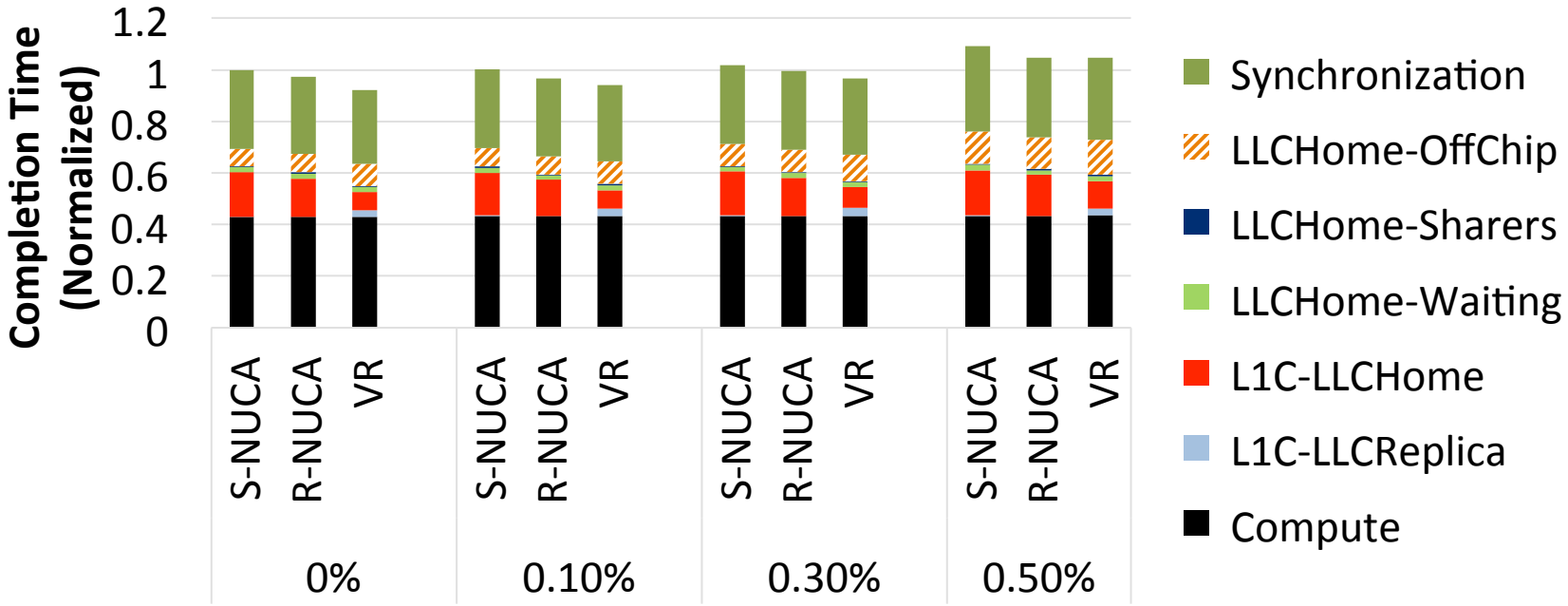
NTV Fault Model for LLC

- Normal distribution of error bits in a cache line with random occurrence probabilities



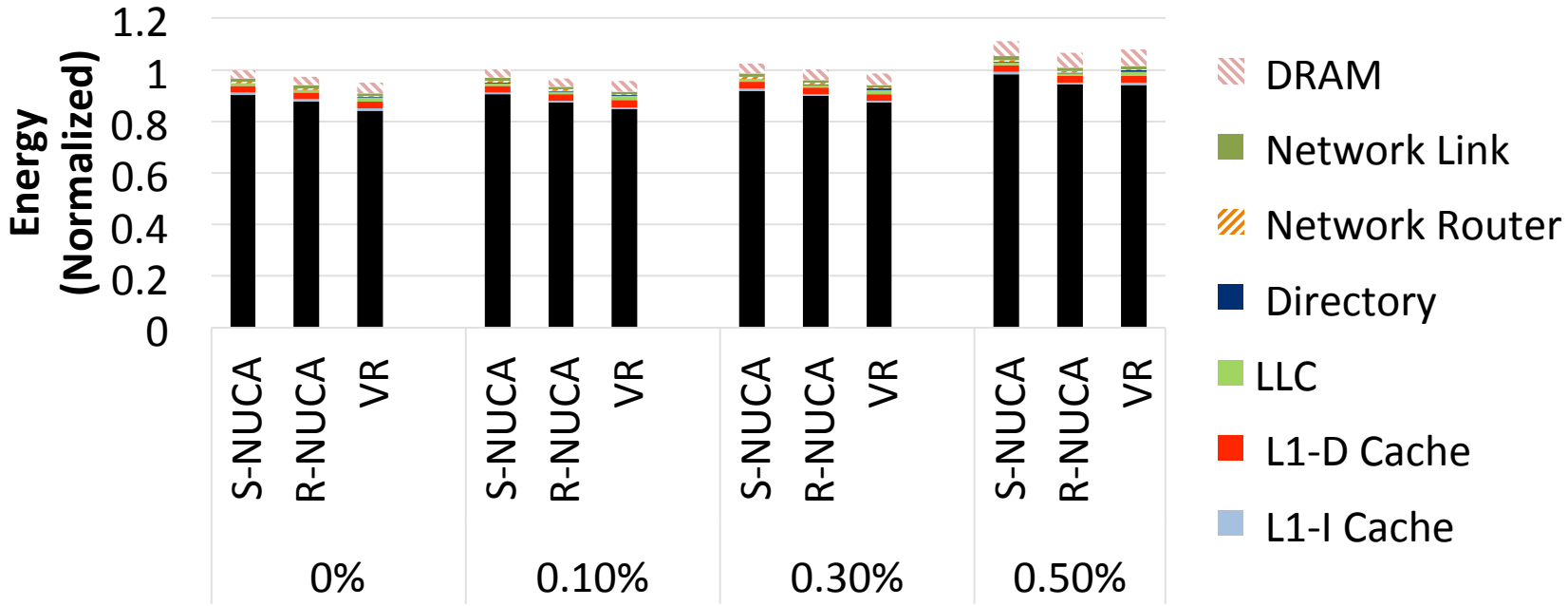
- LLC tag arrays extended to record “disable bits”
  - 0e – 2e: ECC correction with additional 1-cycle latency
  - >2e: Cache line disabling

Average Results – Completion Time



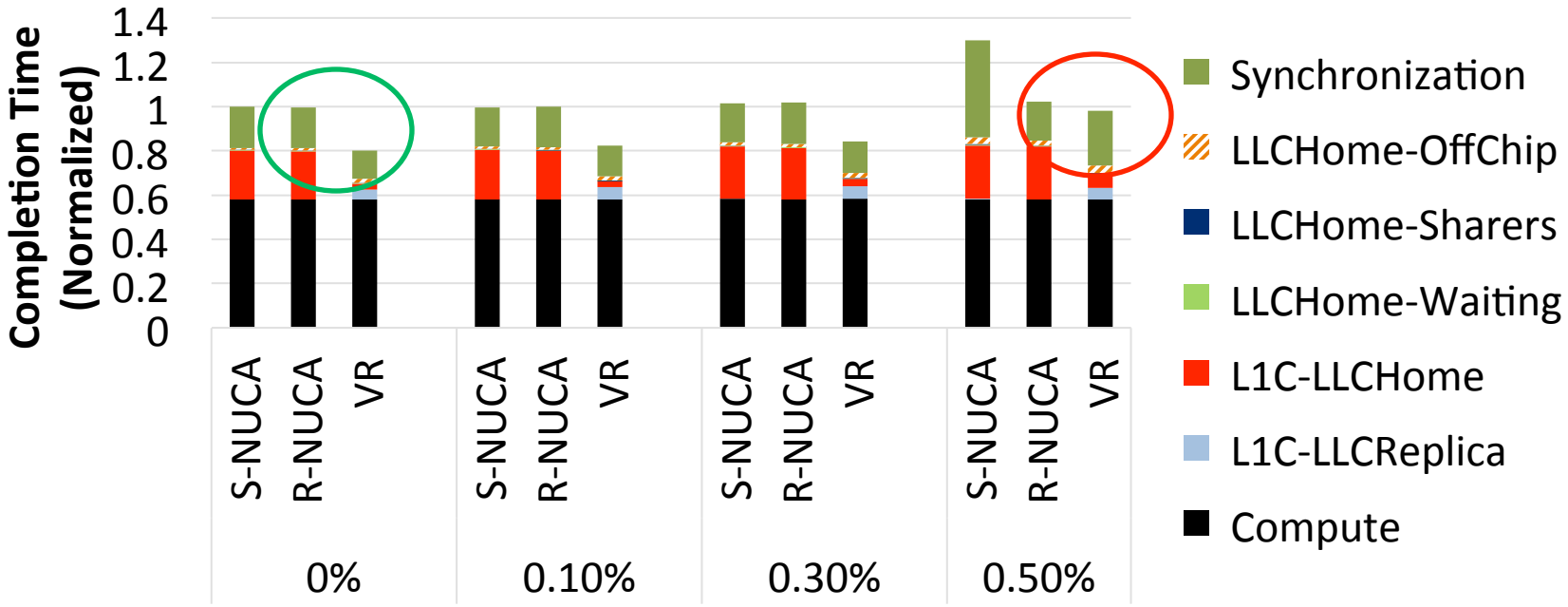
- R-NUCA and VR perform consistently better than S-NUCA
- VR's replication helps at low fault rates
- Lower replication opportunities for VR at higher fault rates result in completion time on-par with R-NUCA

Average Results – Energy



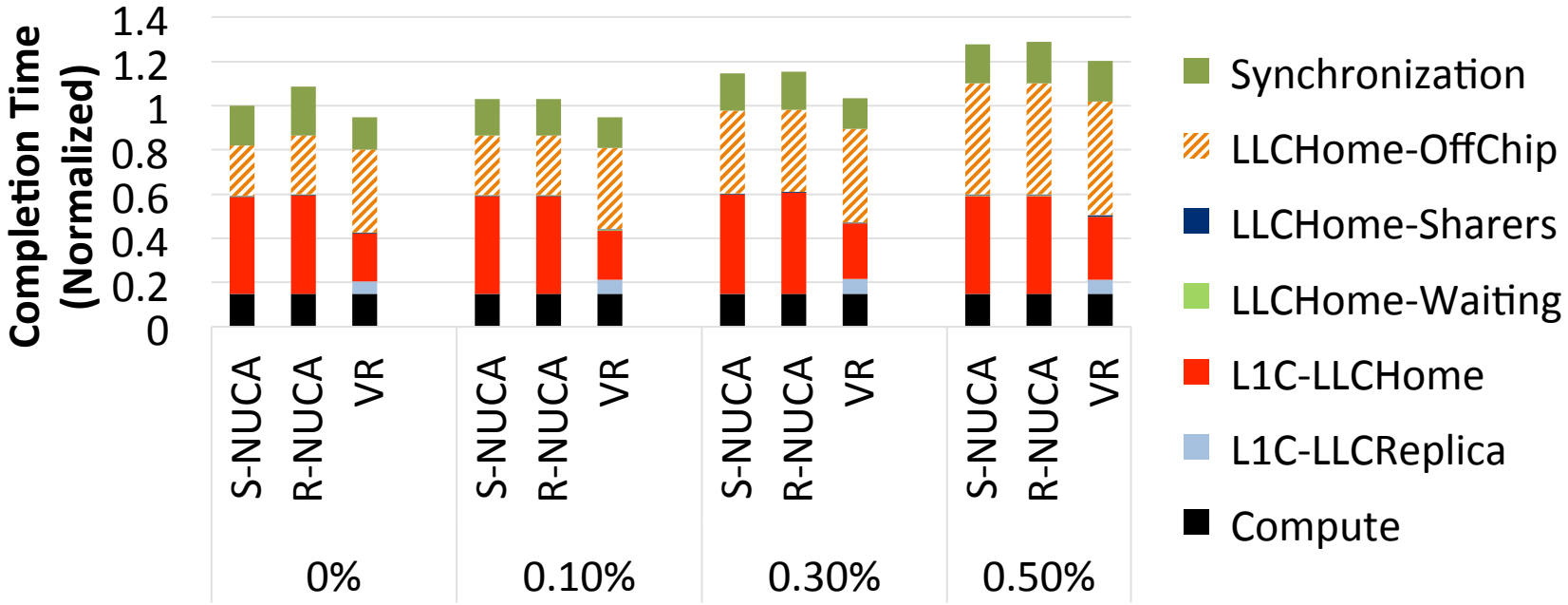
- Static energy dominates the overall energy
- Energy consumption tracks completion time

Benchmark Results – Barnes

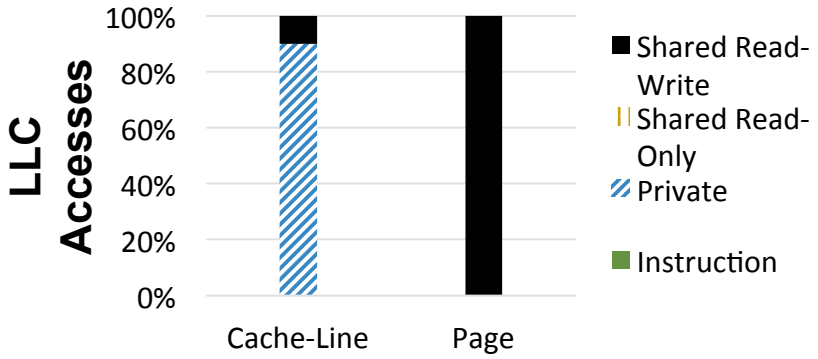


- Replication helps significantly at lower fault rates
- Lower replication opportunity at higher fault rates diminishes advantage over R-NUCA

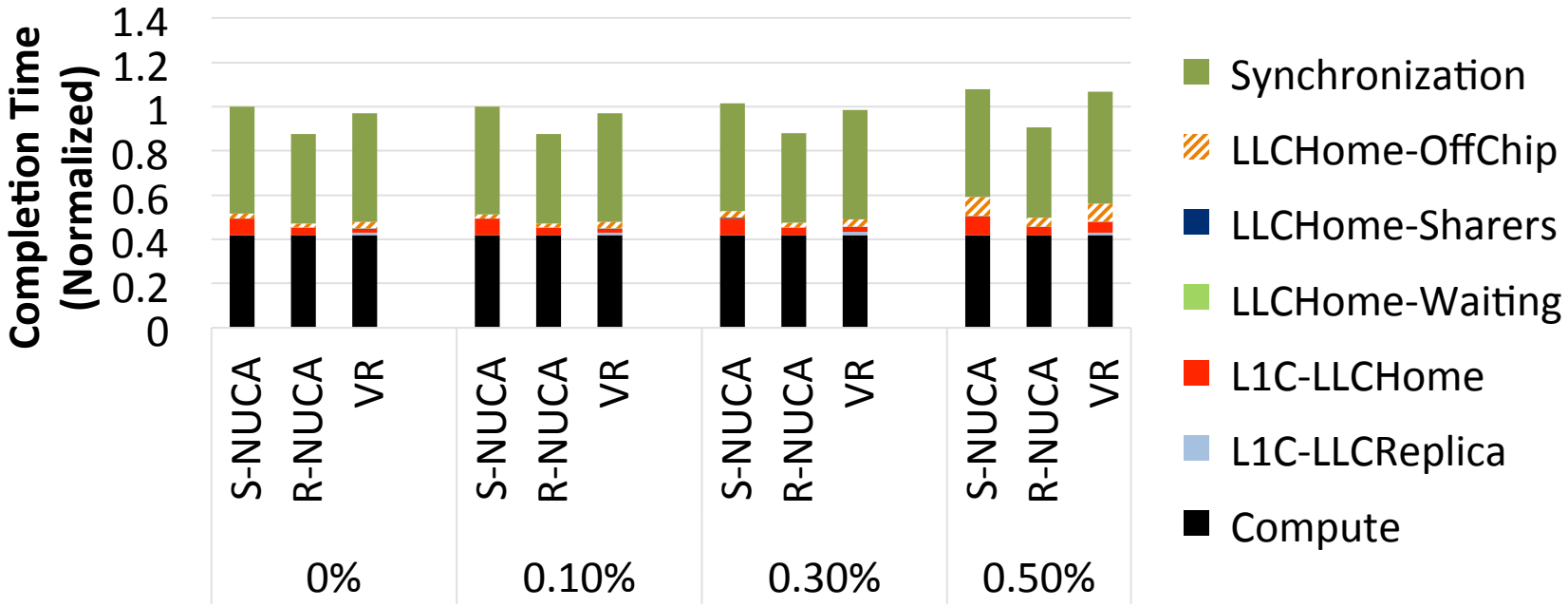
Benchmark Results – Ocean_NC



- R-NUCA performance degrades due to false sharing
- VR better than R-NUCA, however, lower advantage at higher fault rates



Benchmark Results – Dedup



- High number of LLC accesses to thread-private data
- R-NUCA's local placement of private data is effective in improving completion time over VR

Observations

- No one-fits-all data management scheme at the lower LLC capacity when operating at NTV
- A scheme that works optimally at higher LLC capacity might not be effective at the lower usable capacity
- Optimizing locality ends up putting extra stress on the LLC, increasing the off-chip miss rate
  - **There is a need for a data management scheme that not only utilizes LLC capacity more intelligently but also possess the ability to handle the random distribution of faults**