POLITECNICO DI MILANO

DIPARTIMENTO DI
ELETTRONICA,
INFORMAZIONE
E BIOINGEGNERIA

Variation-Aware Voltage Island Formation for Power Efficient Near Threshold Manycore Architectures

ASP-DAC 2014

Stamelakos Ioannis, Sotirios Xydis, Gianluca Palermo,Cristina Silvano
Politecnico di Milano
Dipartimento di Elettronica, Informazione e Bioingegneria

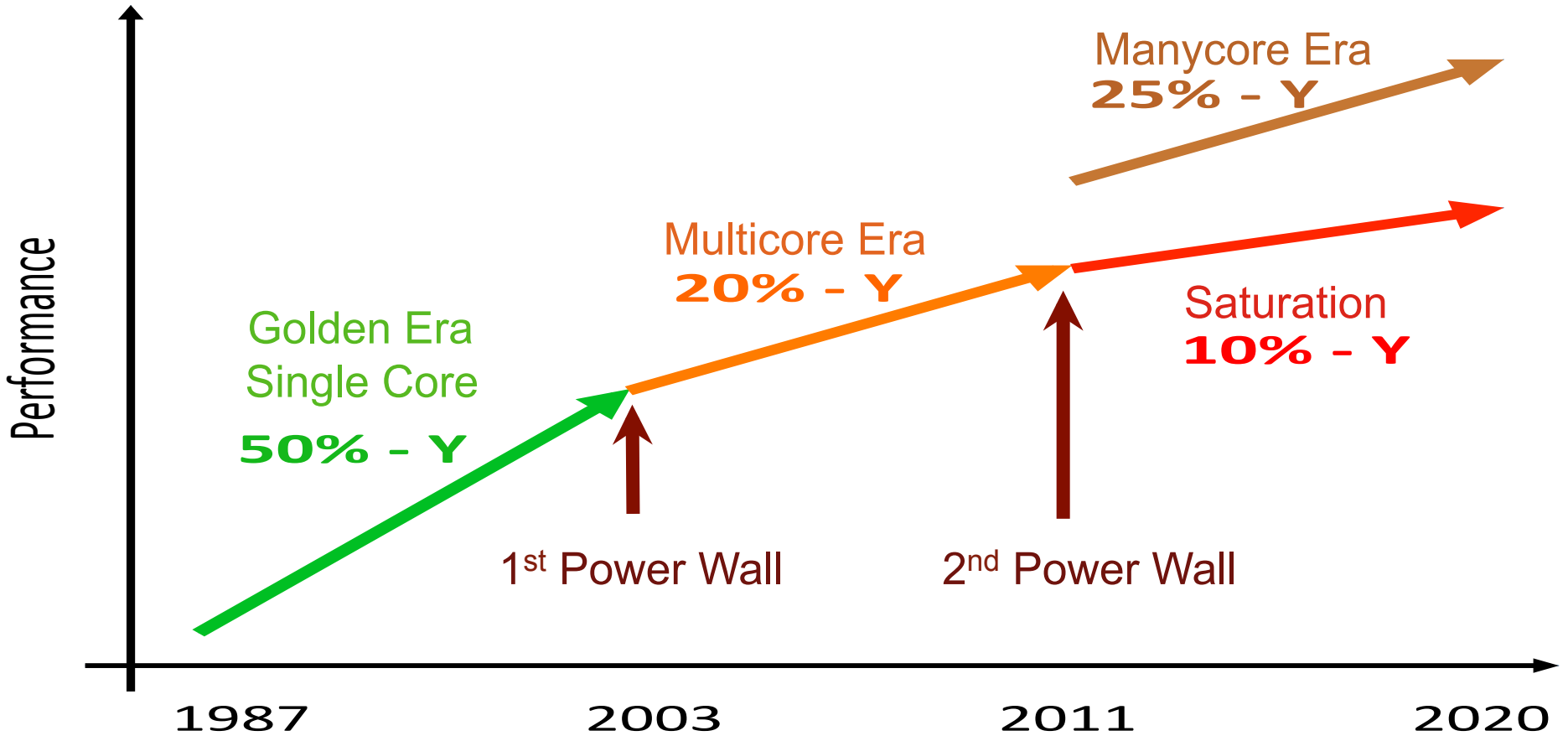
Presenter: ioannis.stamelakos@polimi.it

Outline

- Motivation
- Problem Specification
- Proposed Solution/Framework
- Experimental Setup
- Experimental Results
- Conclusion

Motivation

The Power/Utilization Wall

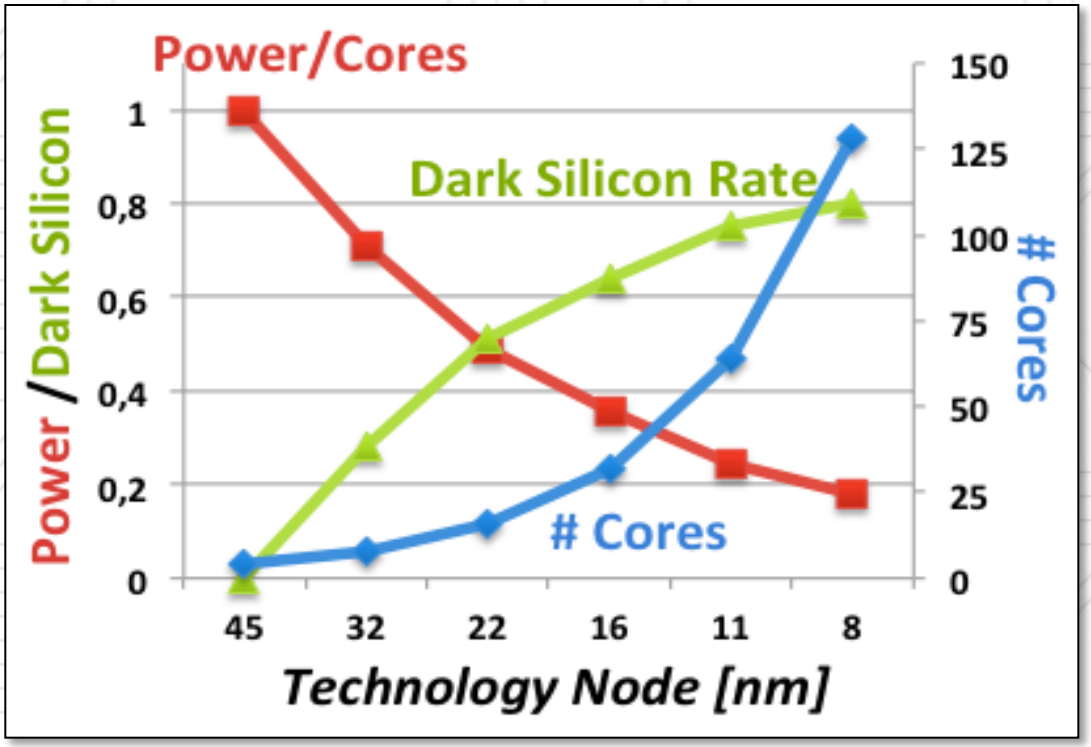


Utilization Wall: The percentage of a chip that can switch at full frequency drops exponentially due to power constraints.

The Dark Silicon Era

Dark Silicon:

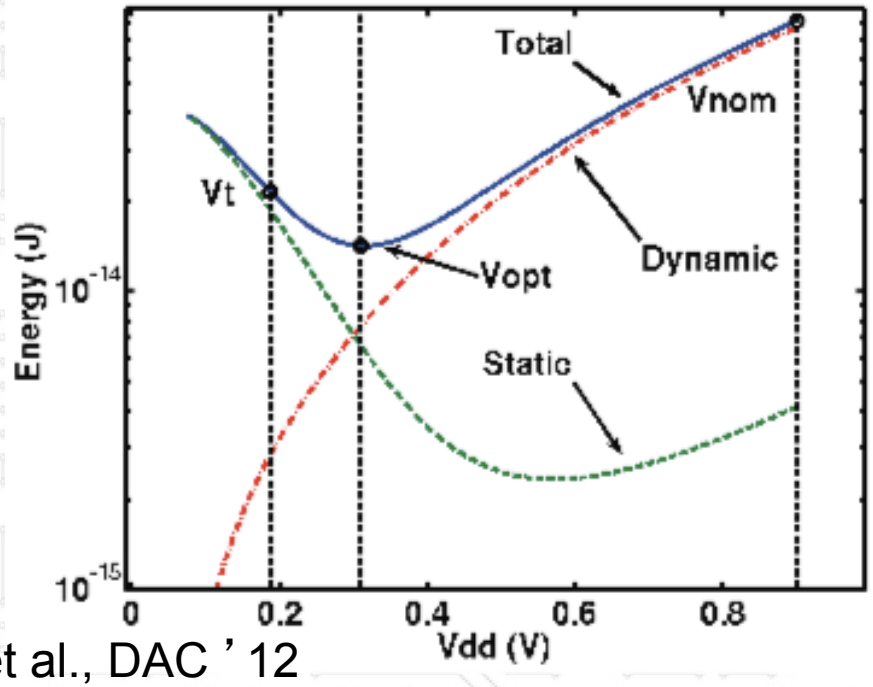
The percentage of transistors/ circuit that is switched off ("dark") due to the limited power budget



Esmaeilzadeh et al., ISCA'11

Near Threshold Computing

- Vdd aggressively tuned close to the Vth value of the transistors
- Lower frequency but larger number of cores available
- Promising energy savings (**10x**) while sustaining performance through parallelization



Pinckney et al., DAC ’ 12

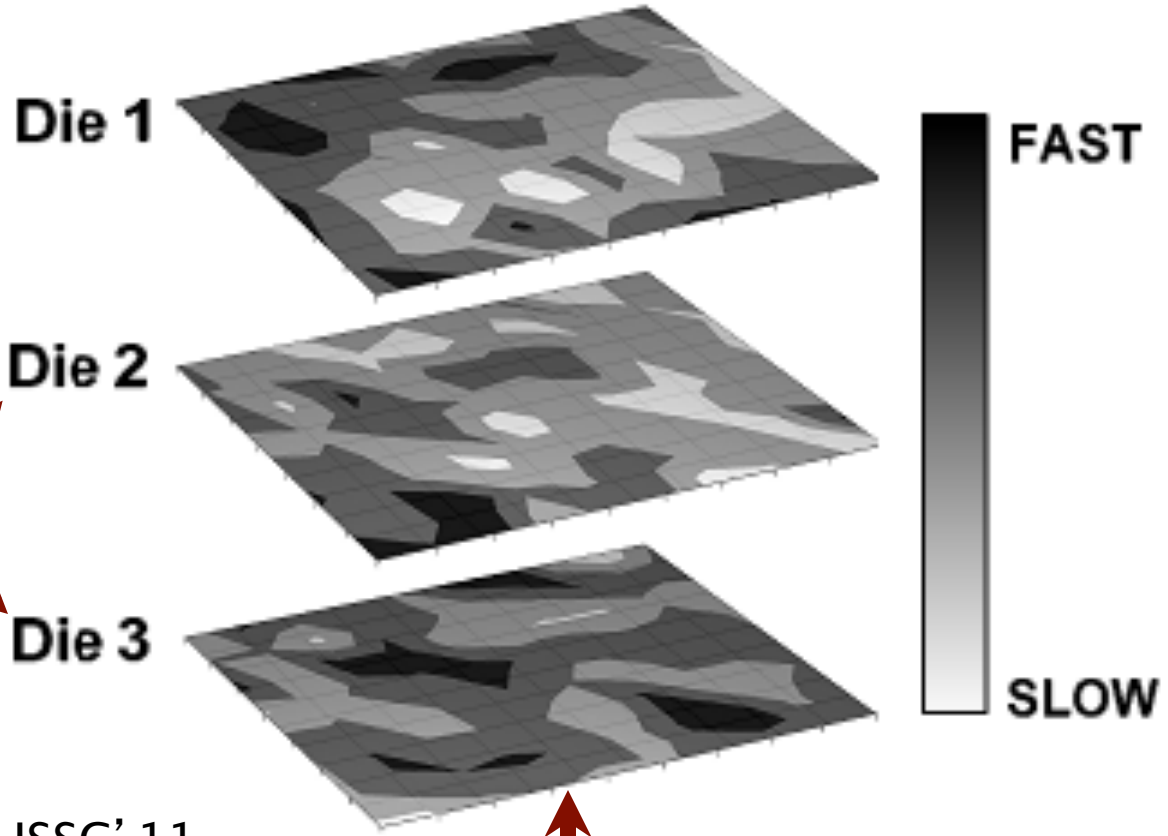
Problem Specification

Performance Degradation @ NTC

- Limited maximum achievable clock frequency
  - Vdd-Vth difference reduction imposes a significant performance degradation
- *Open Issue*: How to sustain performance when exploiting higher task parallelism at lower clock frequencies under process variability ?

The Variability Problem

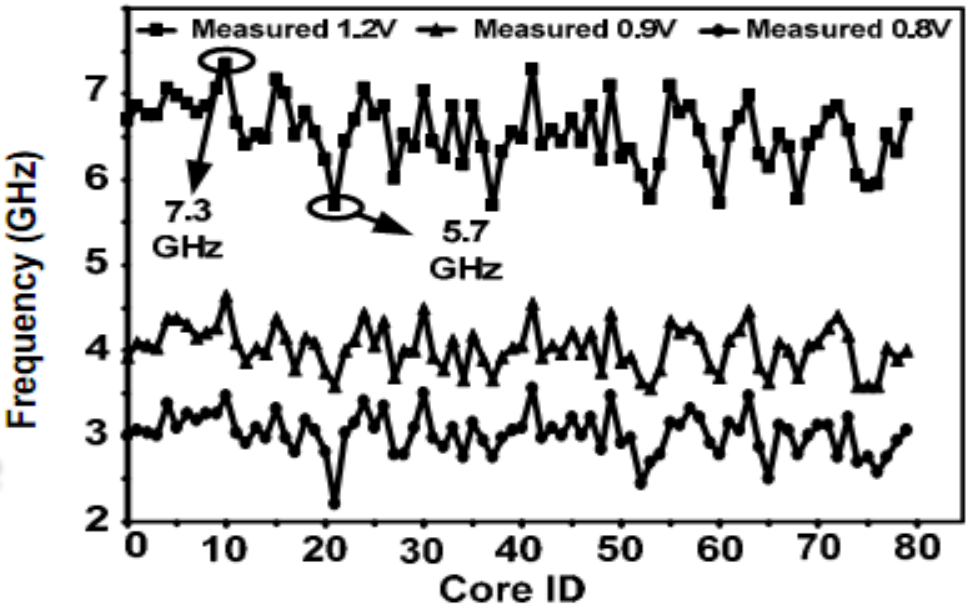
Variability:
the deviation of device parameters (Vth, eff. channel length, etc.) from their nominal values



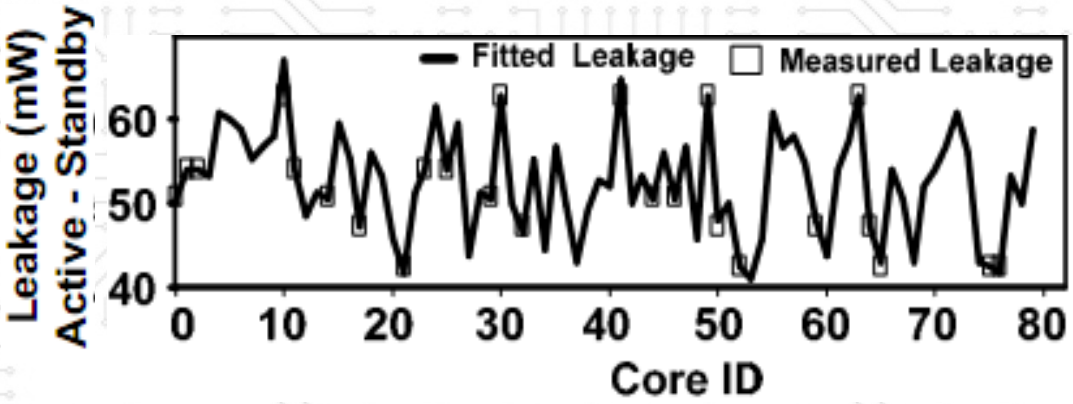
Dighe et al., JSSC' 11

Variability @ NTC

NTC circuits exhibit increased sensitivity to process variations



1.7X deviation in leakage power

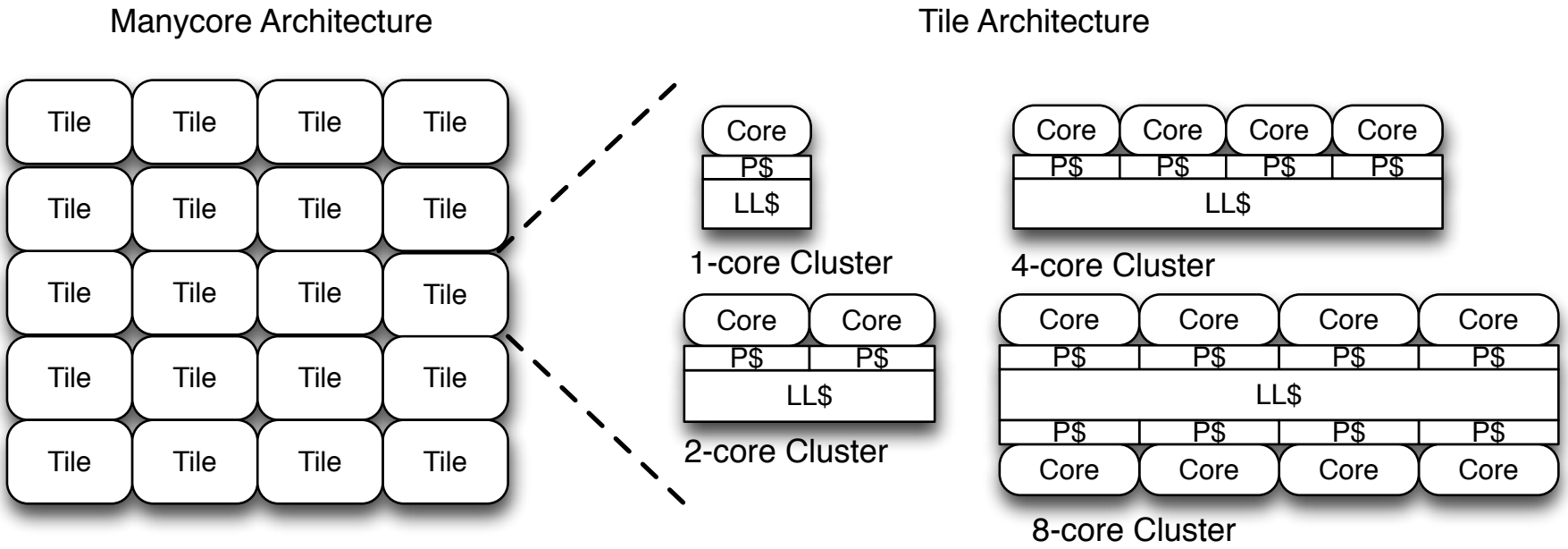


30% deviation in frequency

Dighe et al., JSSC' 11

Proposed Solution/ Framework

Target Architecture

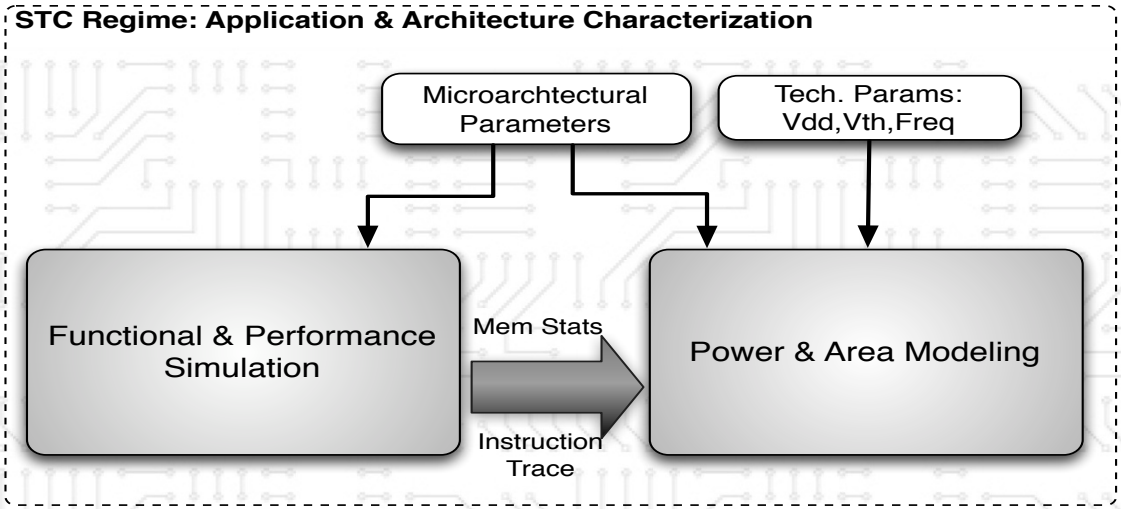


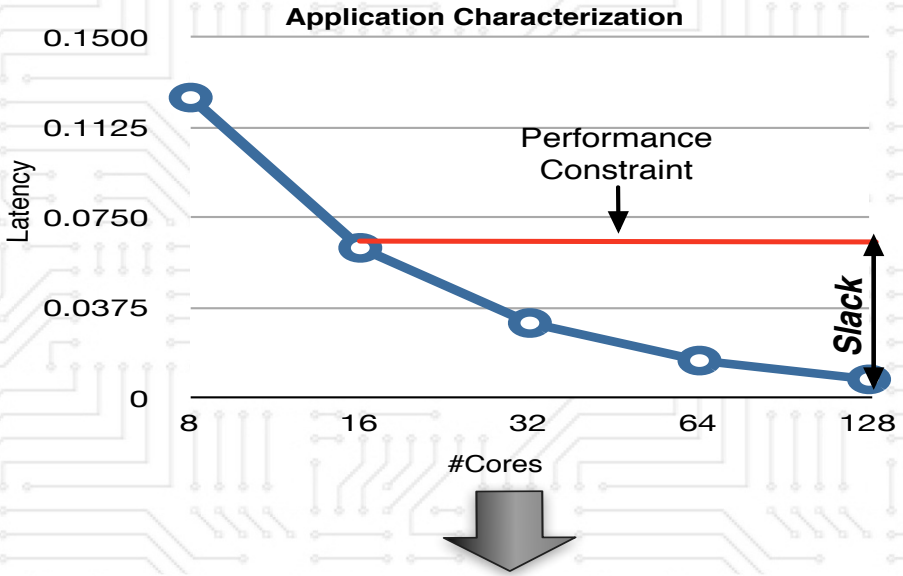
Abstract view of tile based many-core architecture

Voltage Island Formation

- **SFMV**(Single Frequency Multiple Voltages) Approach:
  - One chip-wide frequency but many voltage domains (Voltage Islands)
- Each **VI** can include a certain number of cores and the Vdd can be tuned in a custom way
- Adjust **Vdd** according to the underlying variability in order to reach the desired frequency that sustains the application performance

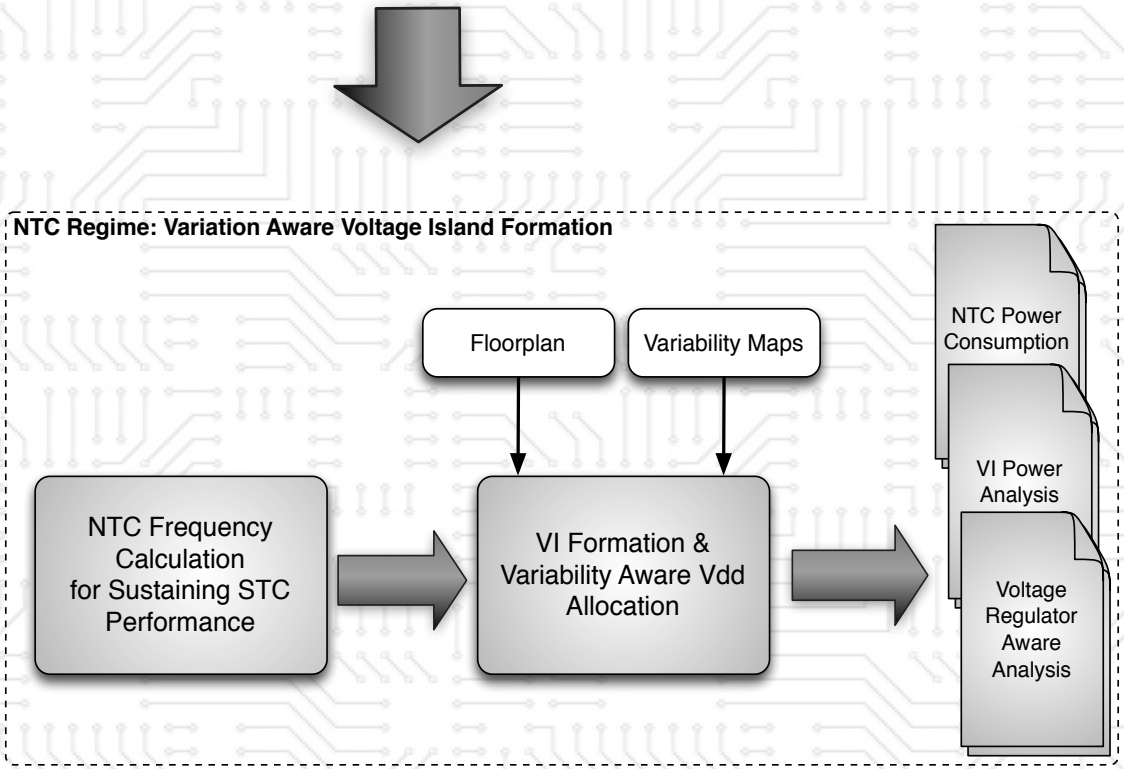
Proposed Framework (I)





Key Point: *Sustain the STC performance at NTC*

Proposed Framework (II)



$$f_{NTC} = \frac{L_{C_{max}}}{L_{min}} \times f_{STC}$$

$$f_{NTC} \propto \frac{(V_{dd}^{(i,j)} - V_{th}^{(i,j)})^{\beta}}{V_{dd}^{(i,j)}}$$

$$V_{dd}^{(k,j)} = \max_{i \in B_k, j \in D} \left[V_{dd}^{(i,j)} \right]$$

Experimental Setup

Experimental Setup

Parameters	Value
Process Technology	22nm
STC Frequency	3.2GHz
STC Supply Voltage	1.05V
Nominal $V_{th}/\sigma_{V_{th}}$	0.23V/0.025
Core Area	$6mm^2$
Private Cache Size/Area	320KB/$4.14mm^2$
Last Level Cache S_i Size – Area	$(2 \times i)$ MB / $(3.88 \times i)mm^2$

Splash-2 Benchmark Suite, run in Sniper sim

Speedup:	ideal	good	limited
	radiosity	*barnes*	*raytrace*
		water-nsq	*water-sp*

+ an average case workload

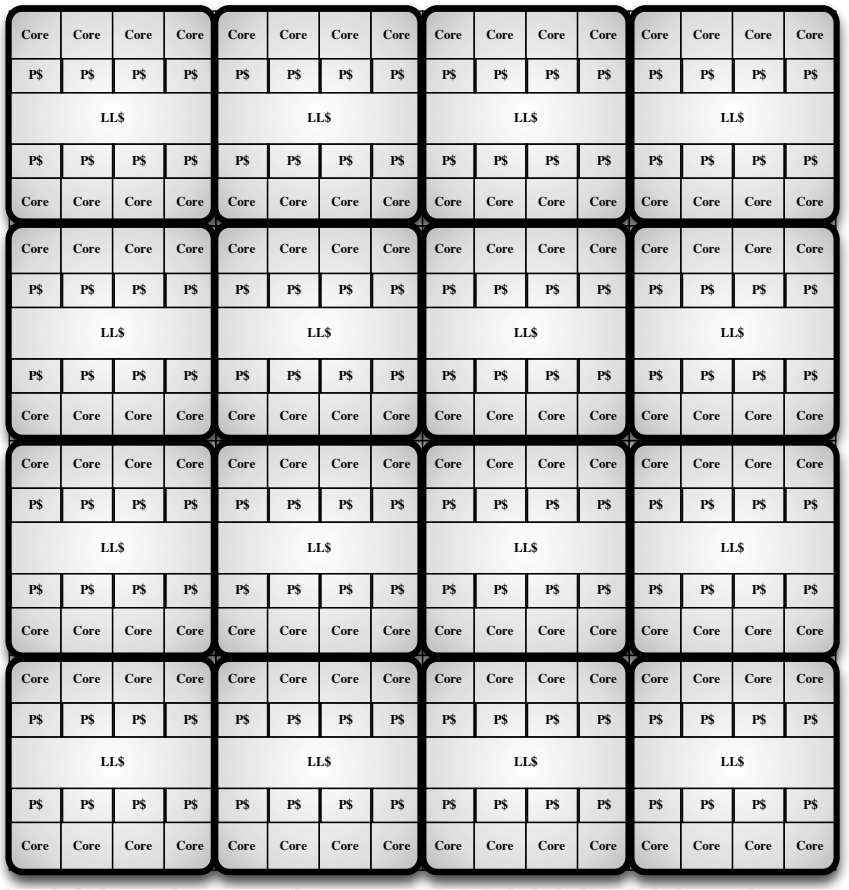
Variability maps: *VARIUS-NTV* Karpuzcu et al., DSN'12

Architecture and Floorplan

A. Tile based many-core architecture

$Tile_{11}$	$Tile_{12}$	$Tile_{13}$	$Tile_{14}$
$Tile_{21}$	$Tile_{22}$	$Tile_{23}$	$Tile_{24}$
$Tile_{31}$	$Tile_{32}$	$Tile_{33}$	$Tile_{34}$
$Tile_{41}$	$Tile_{42}$	$Tile_{43}$	$Tile_{44}$
$Tile_{51}$	$Tile_{52}$	$Tile_{53}$	$Tile_{54}$
$Tile_{61}$	$Tile_{62}$	$Tile_{63}$	$Tile_{64}$
$Tile_{71}$	$Tile_{72}$	P P P P P P P P LL$ LL$	
$Tile_{81}$	$Tile_{82}$	LL$ LL$ P P P P P P P P	

ArchFP*

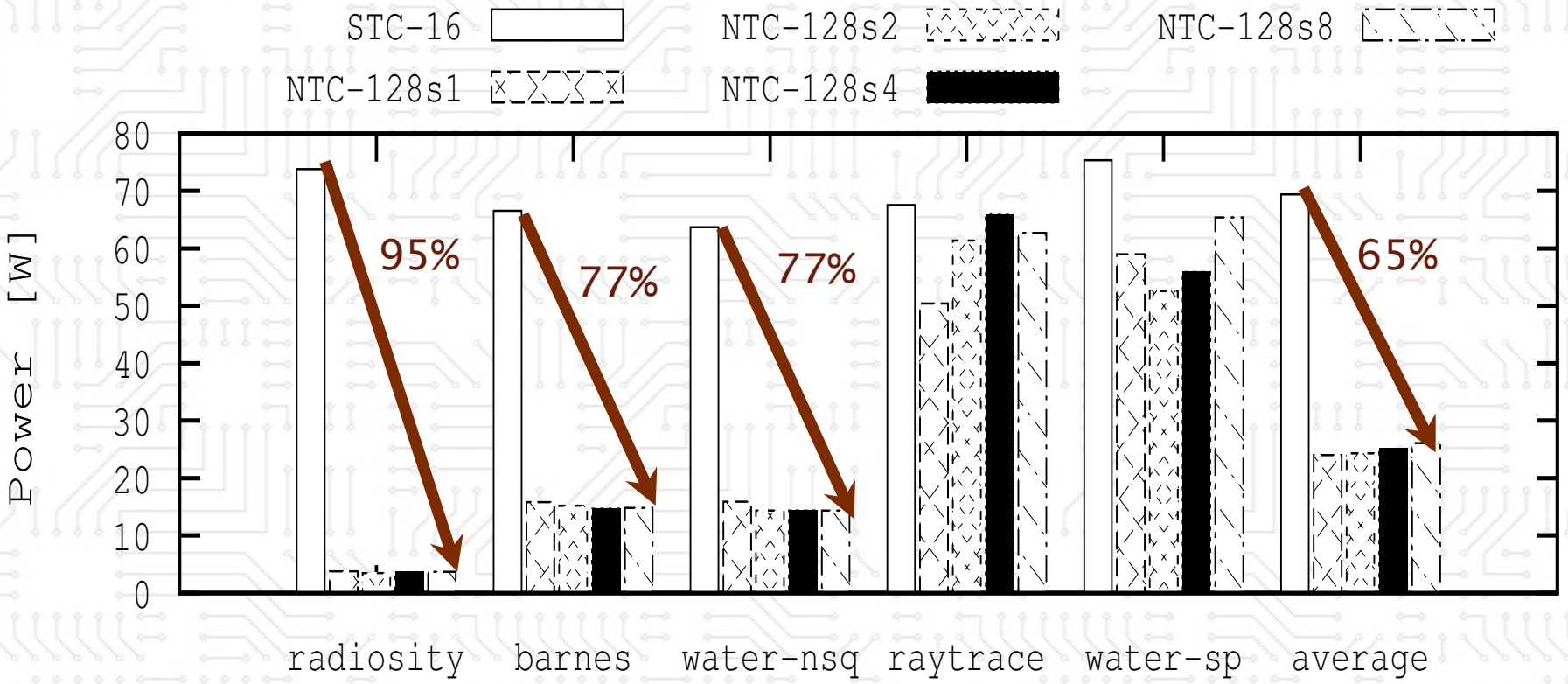


B. Floorplan: 128 cores, 22nm technology

*Faust et al., VLSI-SoC '12

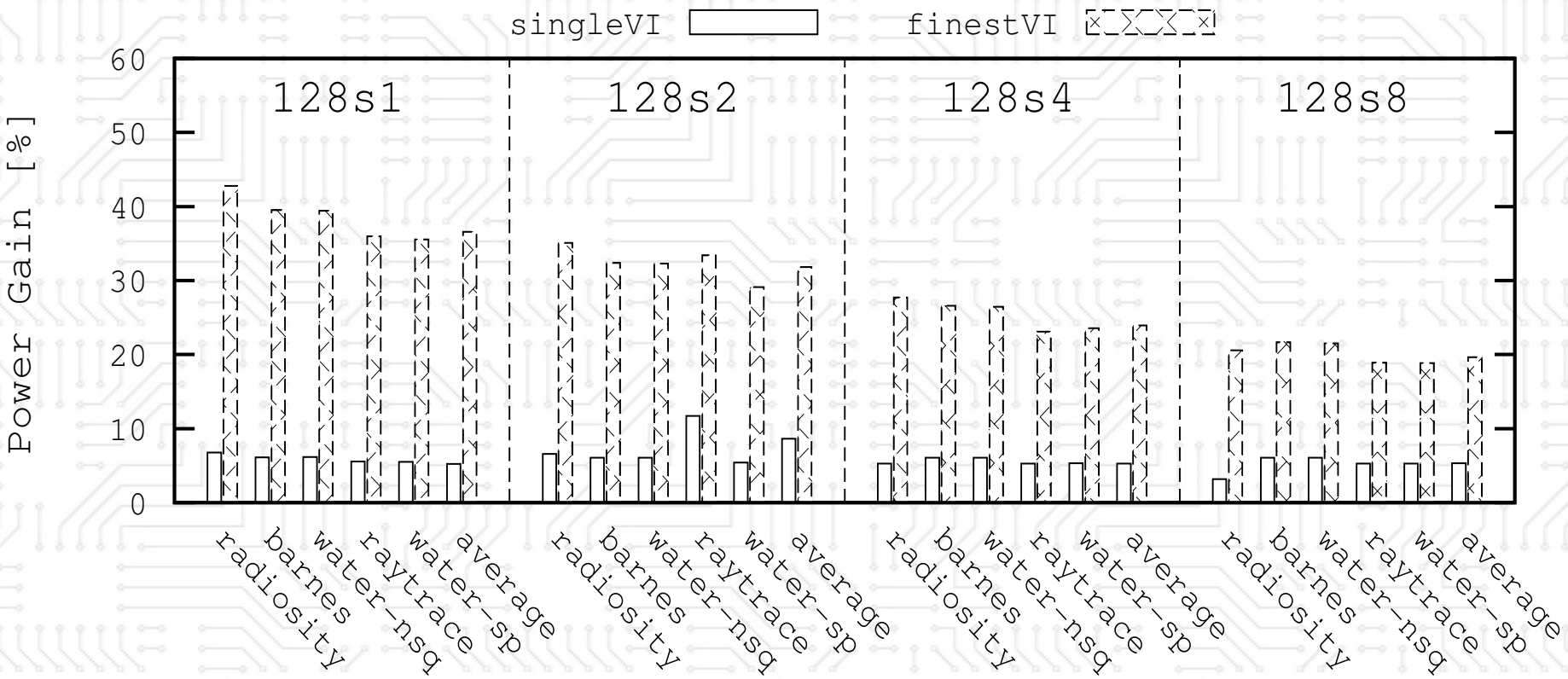
Experimental Results

Power Reduction @ NTC



16-core STC versus 128-core NTC

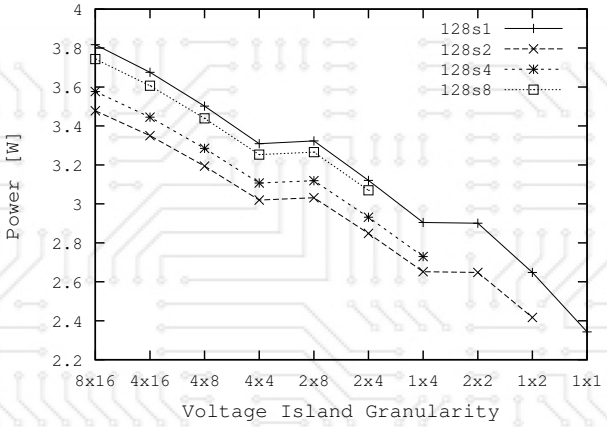
Power Gain of Variability-aware technique w.r.t Overdesign



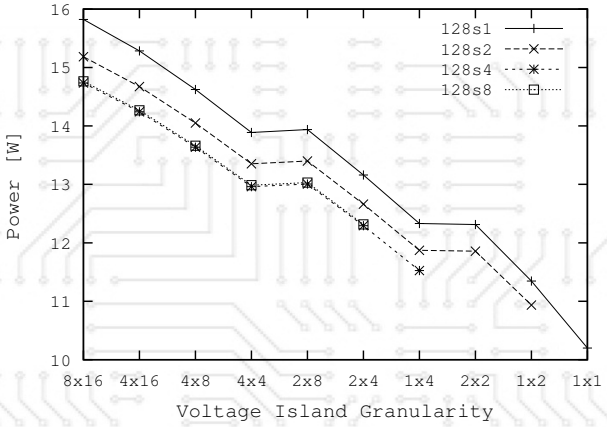
Single VI: ~ 5%

Finest VI:
S1: 34-42% S4: 25-28%
S2: 29-34% S8: 18-23%

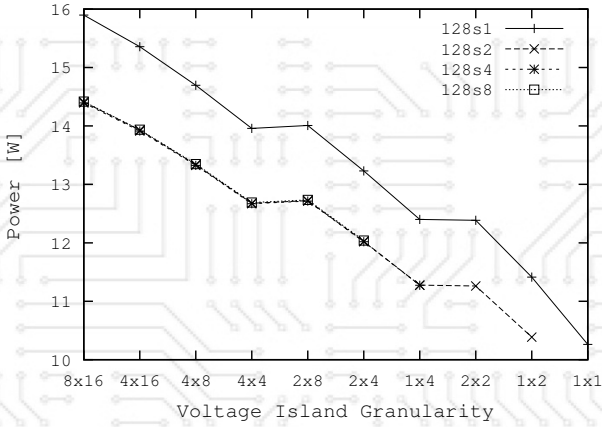
Impact of Voltage Island Granularity on Power Consumption



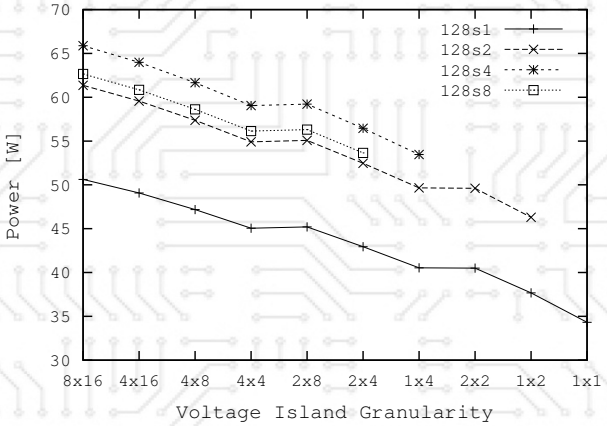
(a) radiocity



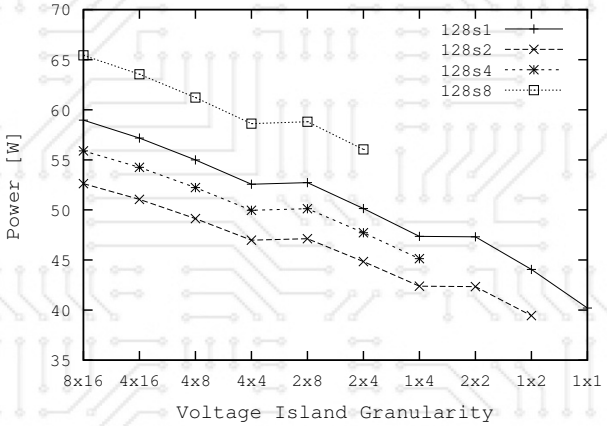
(b) barnes



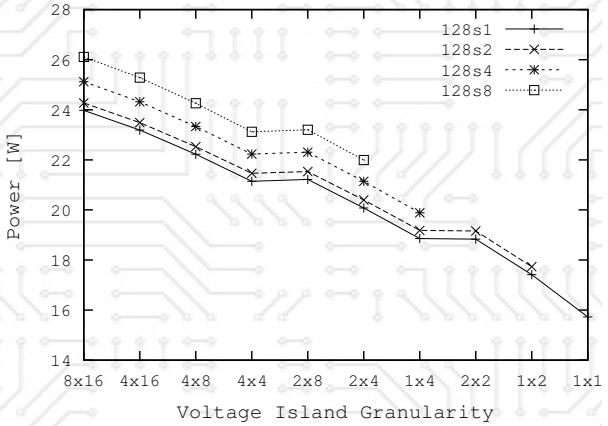
(c) water-nsq



(d) raytrace



(e) water-sp



(f) average

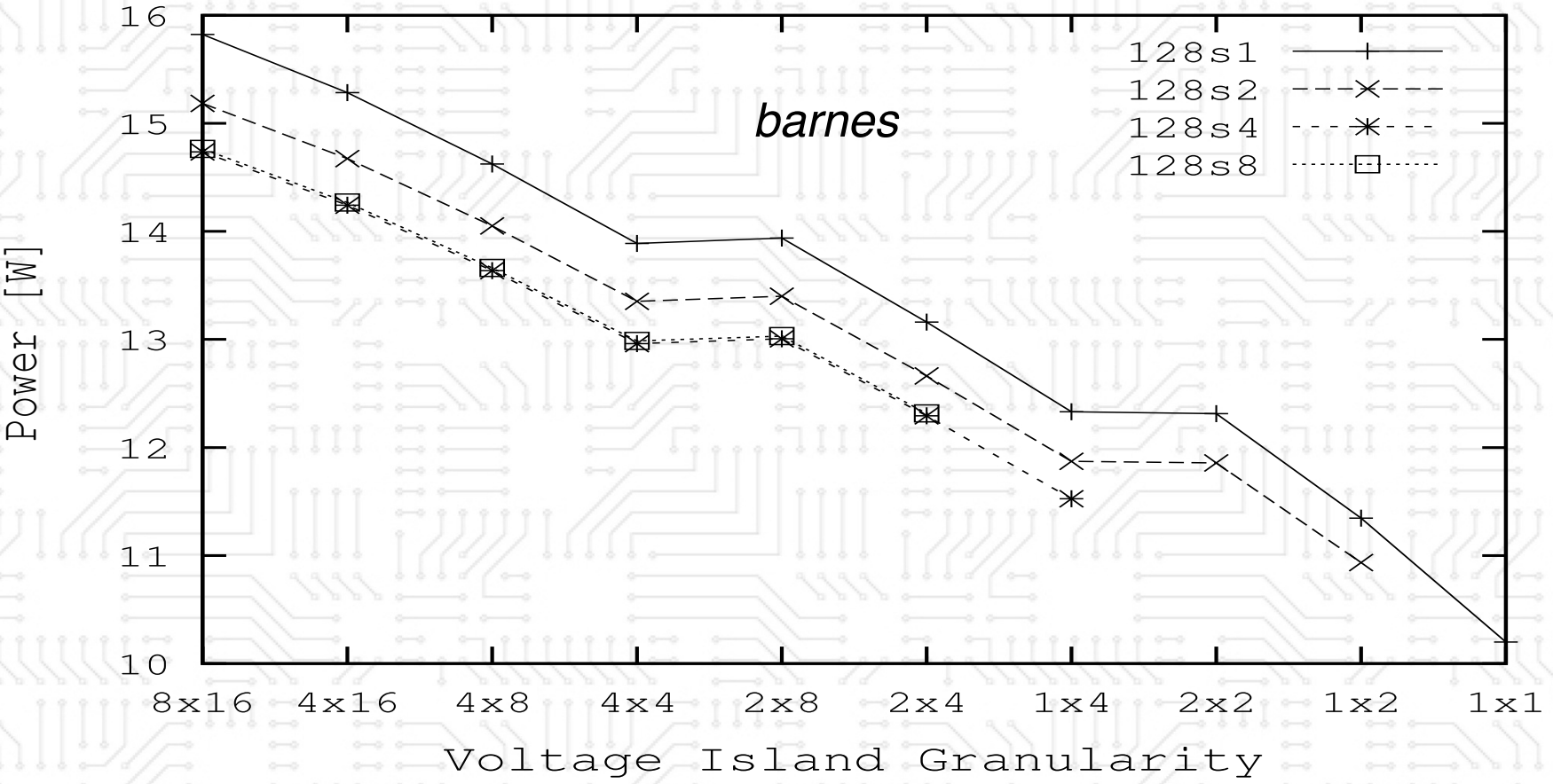
the finer the granularity of the voltage island the higher the power savings.

SingleVI ➔ Finest VI

S1: 30-35% S4: 19-24%

S2: 24-30% S8: 14-18%

Impact of Voltage Island Granularity on Power Consumption



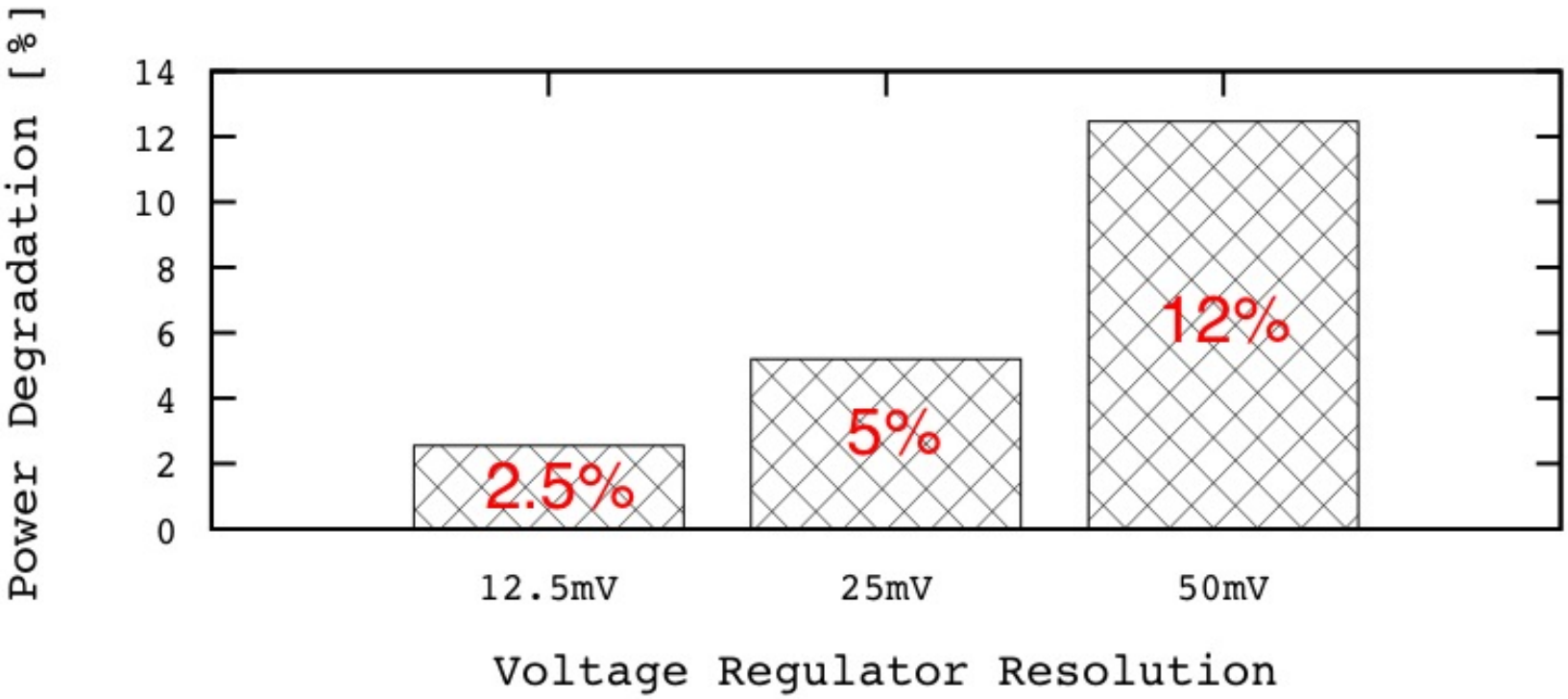
the finer the granularity of the voltage island the higher the power savings.

SingleVI ➔ Finest VI

S1: 35% S4: 21%

S2: 28% S8: 16%

Impact of Voltage Regulator Resolution on power efficiency at NTC



- Power overhead: the normalized difference between the power consumed in the ideal case and the power with the specific value of voltage precision
- The higher is the resolution the smaller is the overhead
- Even the 12% can be tolerable for applications that exhibit ideal or good scaling

Conclusion

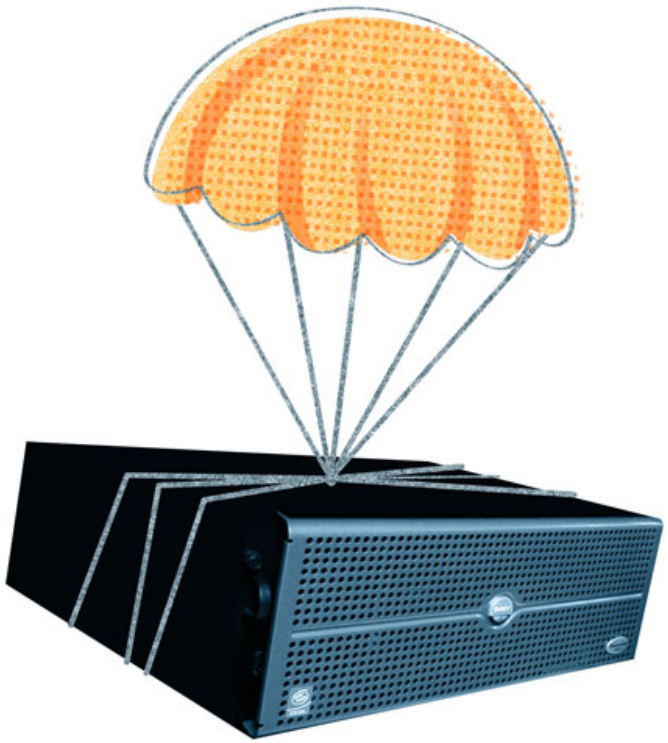
Conclusions

- A variability-aware framework for exploring the power-efficiency of Near-Threshold Computing
- Voltage island formation combined with the operation at the near-threshold regime proposed as an effective technique for building power efficient many-core architectures while sustaining super threshold performance
- Promising results shown, depending on both workload characteristics and the underlying architectural organization
  - ~ 65% average power gain
  - ~ 15-35% extra savings for finest VI granularity
  - ~ 2.5 -12% power degradation due to VR quantization

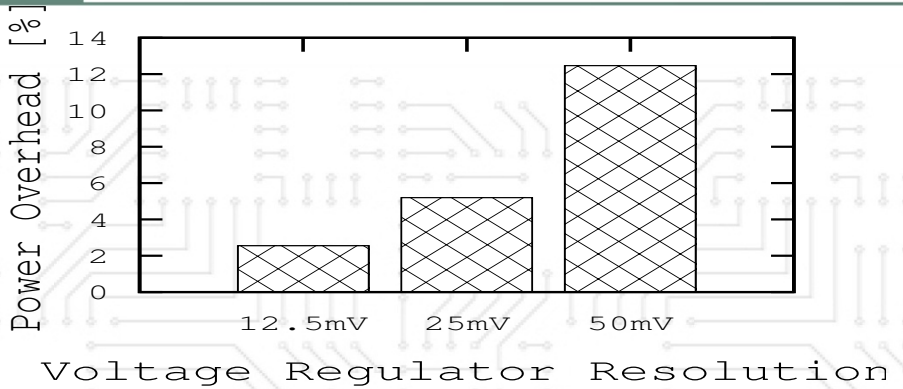
Thank You!!!

Questions???

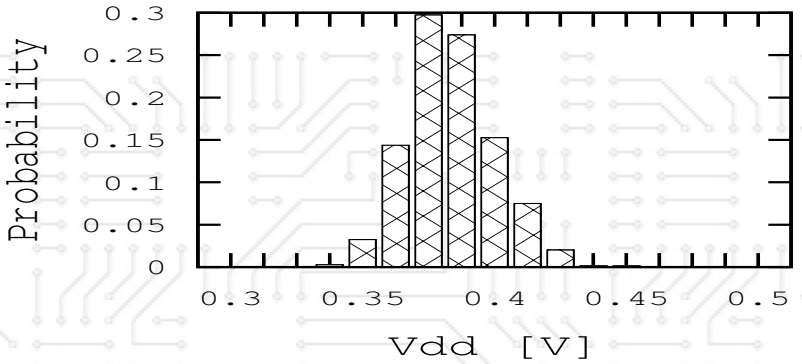
Backup Slides



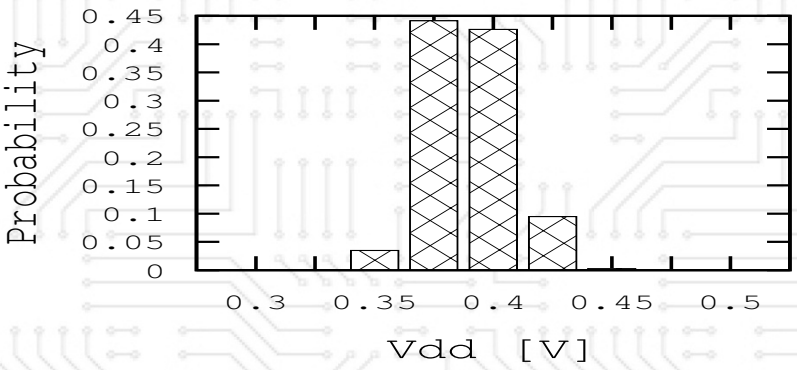
Vdd Distribution at NT regime

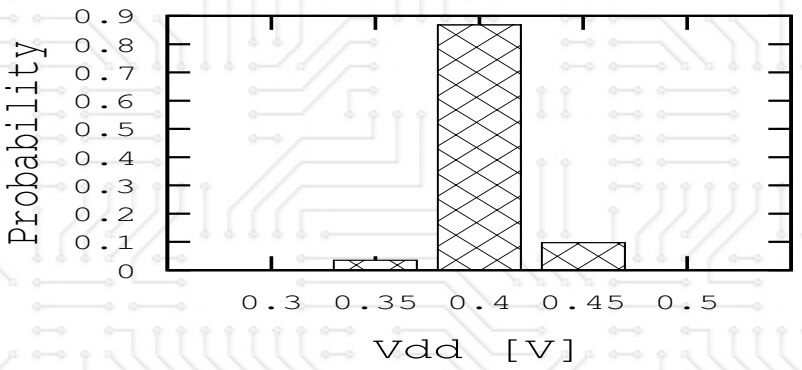


(a) Overhead



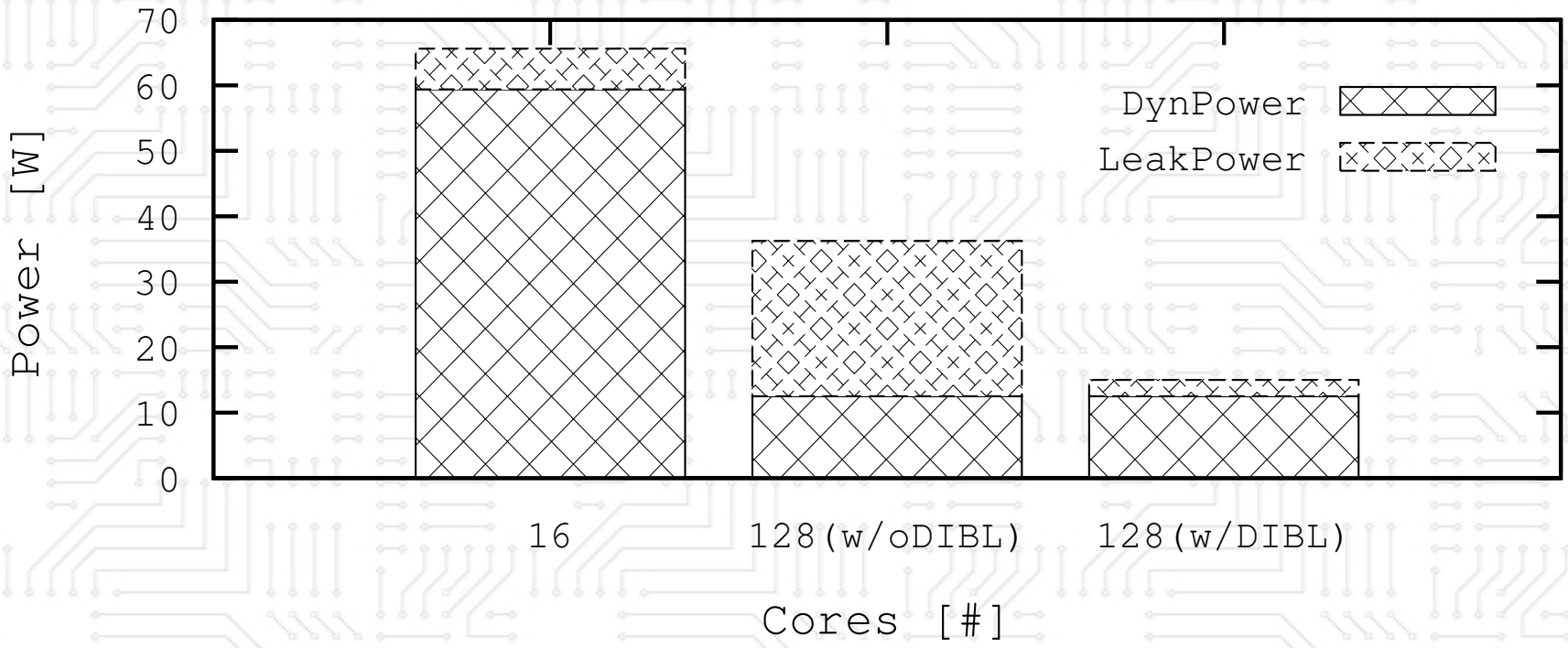
(b) 12.5mV





- Narrow Vdd distribution
- No need of allocating multiple levels of supply voltage

The DIBL Effect



$$I_{sub} = I_0 e^{\frac{V_G - V_S - V_{T0} - \gamma V_S + \eta V_{DS}}{n V_{th}}} \left(1 - e^{\frac{-V_{DS}}{V_{th}}}\right)$$

Vth = thermal voltage

n = subthreshold swing coefficient

γ = linearized body effect coefficient,

η = DIBL coefficient