www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Fault Characterization Through FPGAs Undervolting

Behzad Salami, Osman S. Unsal, and
Adrian Cristal Kestelman

28th Field Programmable Logic & Applications (FPL) Conference,
27-Aug-2018, Dublin, Ireland.

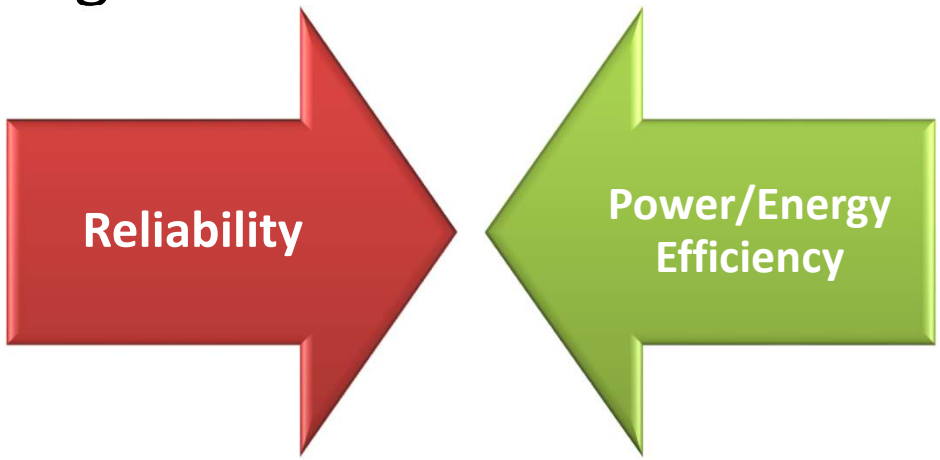



Undervolting

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Underscaling the supply voltage *below the nominal level* :

- **Power/Energy Efficiency**: Reduces quadratic ally dynamic and linearly static power.
- **Reliability**: Increases the circuit delay and in turn, causes timing faults.



Aggressive Undervolting is not DVFS!

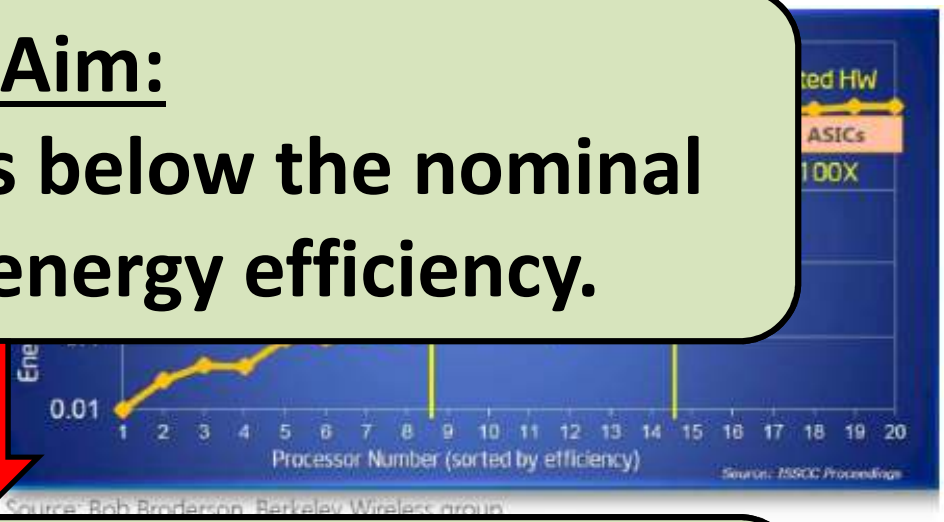
Motivation

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Contribution of FPGAs in large data centers is growing, expected to be in 30% of datacenter servers by 2020 (Top500 news).

Our Aim:

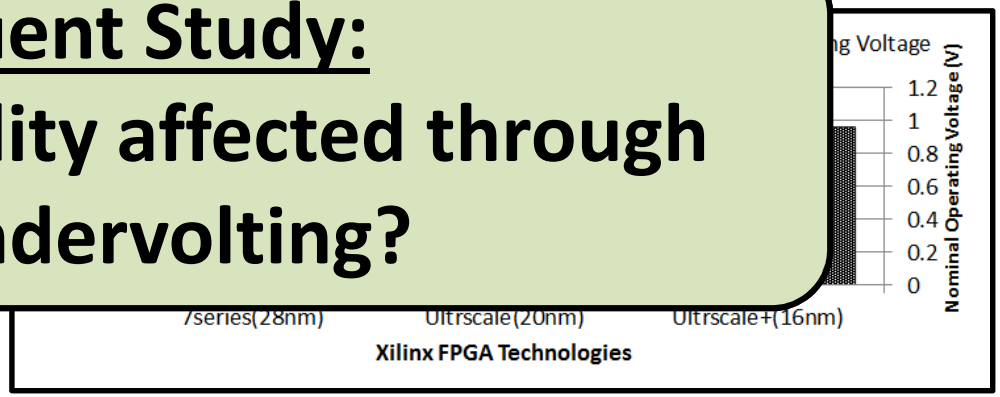
Undervolting FPGAs below the nominal level to achieve energy efficiency.



Source: Bob Broderson, Berkeley Wireless group

Subsequent Study:

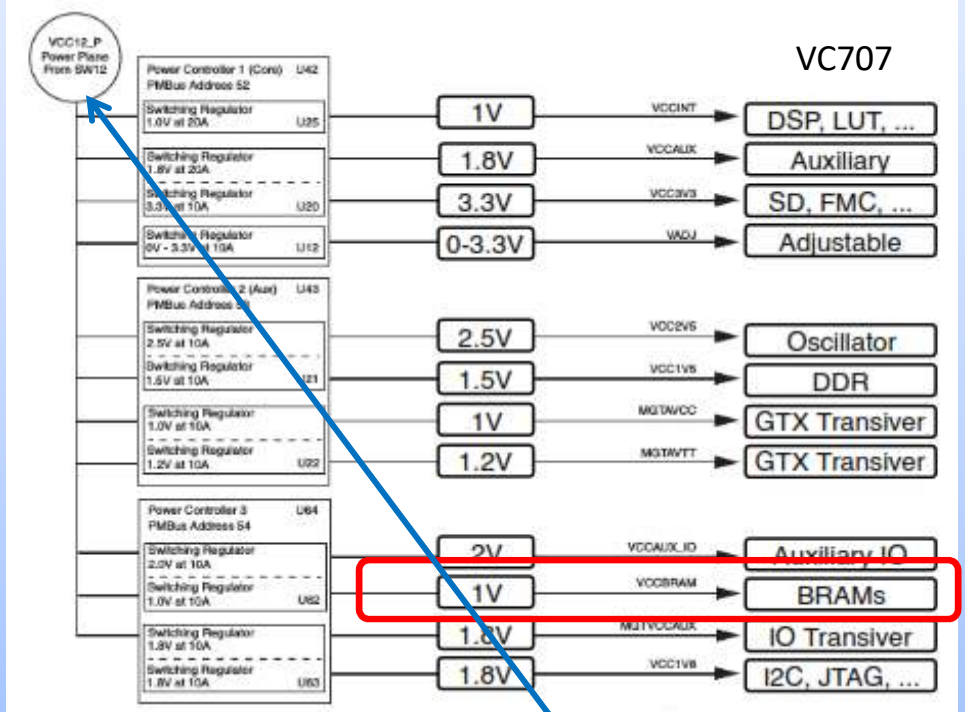
How is the reliability affected through FPGAs Undervolting?



Voltage Scaling Capability in Xilinx

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Voltage Distribution on Xilinx Platforms



Evaluated Xilinx Platforms

VC707: performance-efficient design

KC705: power-efficient design

Voltage Regulator

- Power Management Bus (PMBus).
- Hardwired to the host.

Influencing the Direction of Power Management

PMBus® A·V·S

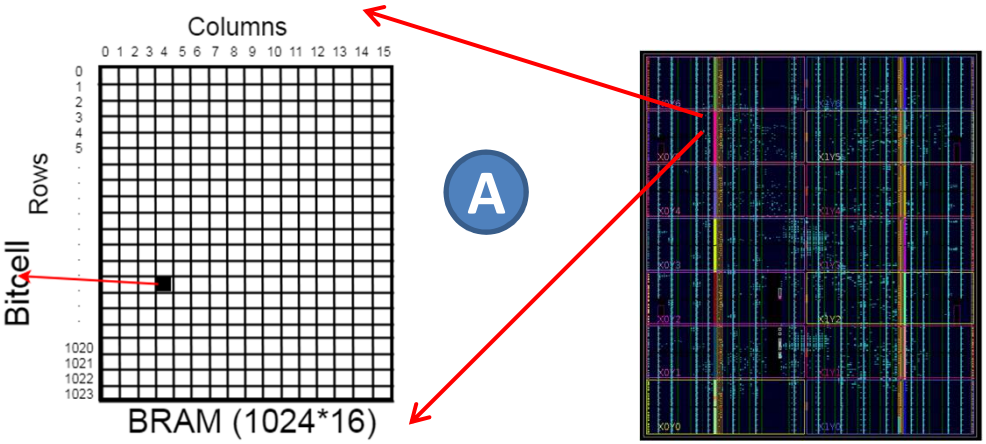
Experimental Methodology

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

A Detailed study on FPGA BRAMs, which are a set of bitcells in the row-column format.

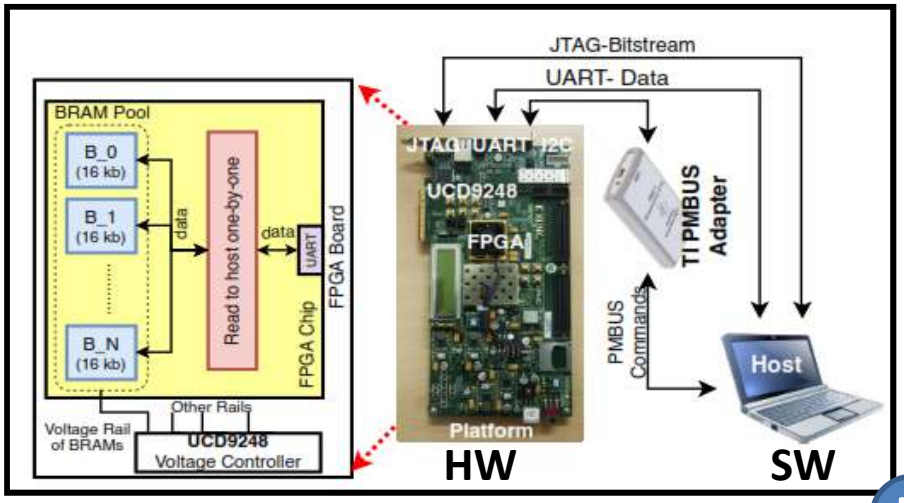
B Experimental Methodology:

1. **HW**: Transfer content of BRAMs to the host.
2. **SW**: Analyze data, and adjust voltage of BRAMs.



Floorplan of VC707

Operating frequency is set to the maximum, i.e., ~500mhz.




```
1: V_CCBRAM = V_min;
2: while(V_CCBRAM >= V_crash) begin
3:     while(numRun <= 100) begin
4:         delay(1sec);
5:         Transfer content of BRAMs to the host;
6:         Analyse faulty data (rate and location);
7:         numRun++;
8:     end
9:     V_CCBRAM− = 10(mV);
10: end
```

B

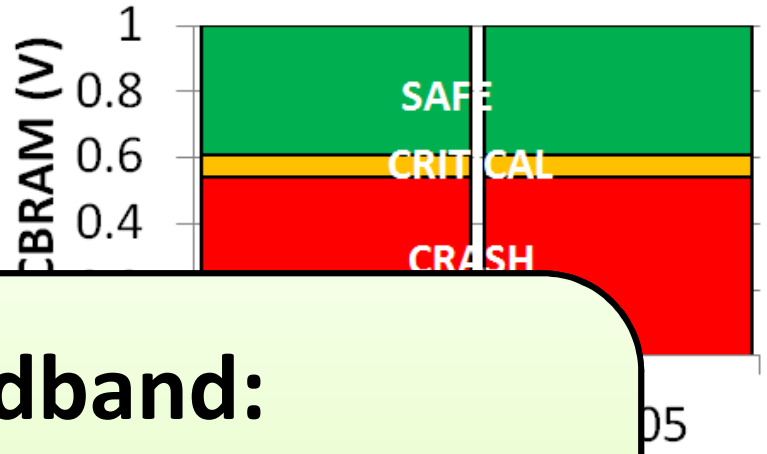
Overall Behavior- Power & Reliability

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

SAFE
- No observable fault
- Voltage Guardband Below Vnom

CRITICAL
- Faults manifest
- Below Vmin, min safe voltage

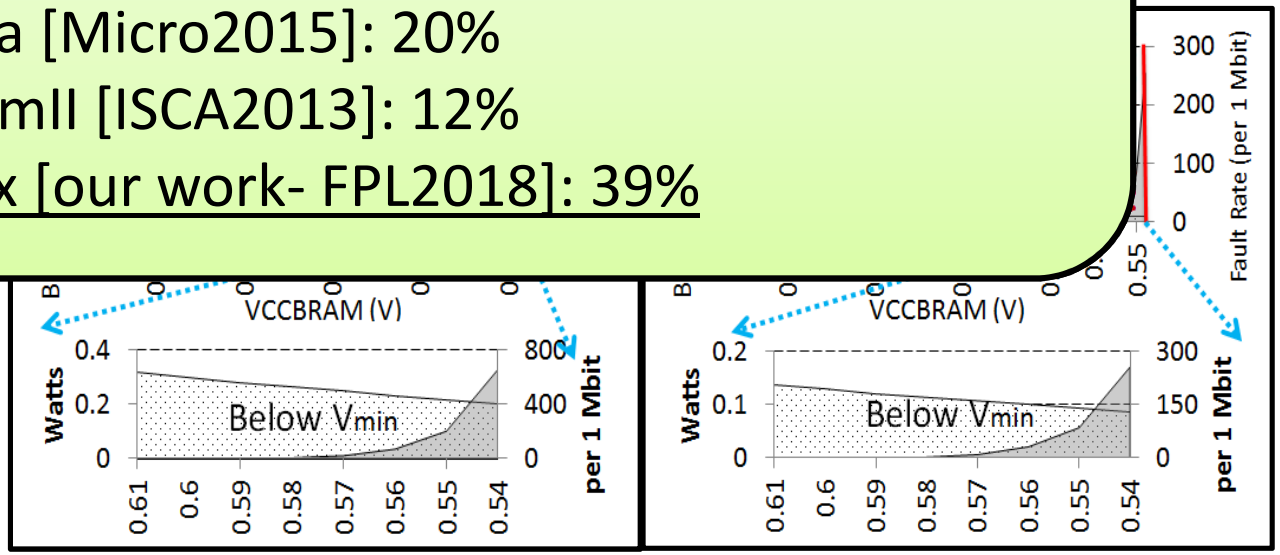
C



Voltage Guardband:

1- DRAM- MultipleVendors [Sigmetrics2017]:16%
2- GPU- NVidia [Micro2015]: 20%
3- CPU- ItaniumII [ISCA2013]: 12%
4- FPGA- Xilinx [our work- FPL2018]: 39%

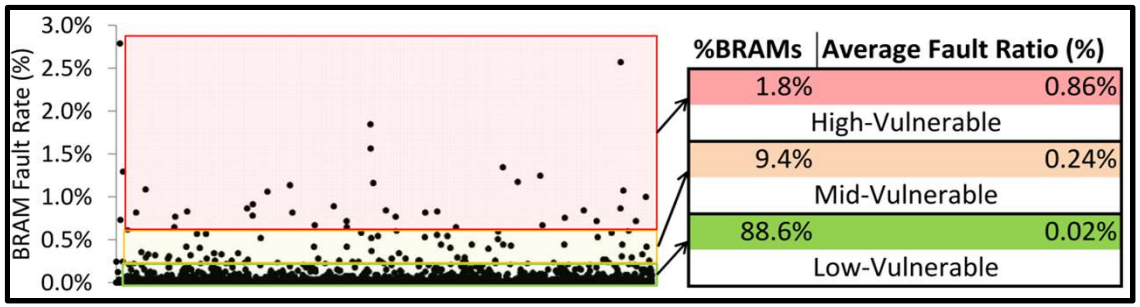
1. V
2. V
   s
3. N
   ef
4. Exponential fault rate increase.
5. VC707 experiences relatively more fault rate.



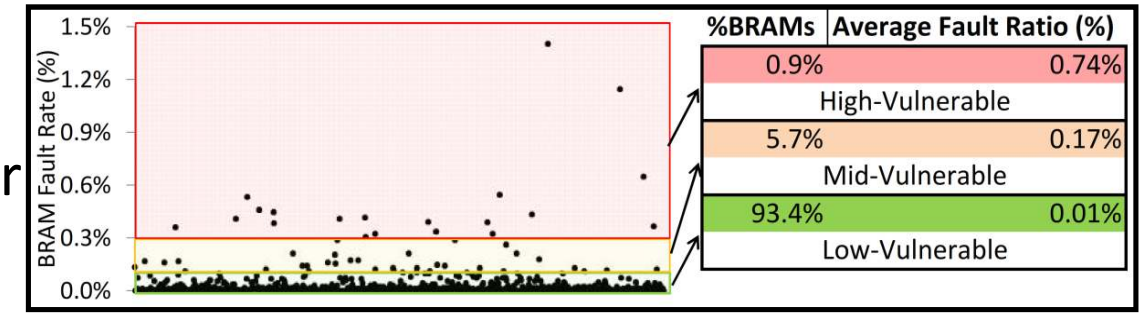
VC707 **KC705**

Fault Variability between BRAMs

- BRAMs clustering using K-Mean clustering.
- Majority of BRAMs are low-vulnerable.
- ~36% of BRAMs never experience faults.
- Fully non-uniform fault distribution.



VC707



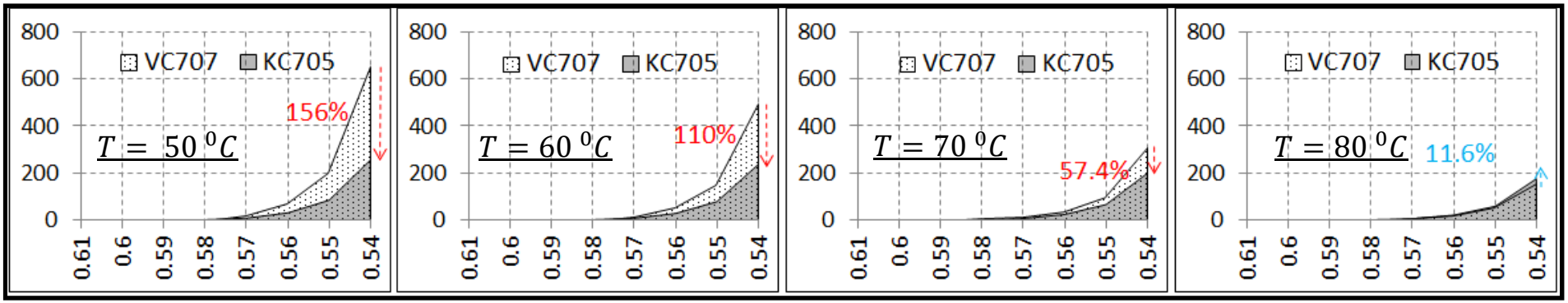
KC705

VCCBRAM= Vcrash

* Different scales in y-axis * *Pattern= 18'h3FFFF *

Environmental Temperature

- **Methodology:** Adjusting environmental temperature, monitoring on-board temperature via PMBus.
- **Experimental Observation:**
  - At higher temperatures, fault rate is significantly reduced.
  - The rate of this reduction is highly platform-dependent (VC707 > KC705).
- **Inverse Temperature Dependency (ITD):**
  - For nano-scale technologies, under ultra low-voltage operations, the circuit delay reduces at higher temperatures since supply voltage approaches the threshold voltage.



* y-axis: VCCBRAM (V), y-axis: fault rate (per 1Mbit) *

Summary & Future Works

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Summary

- We experimentally showed how Xilinx FPGAs work under aggressive low-voltage operations.
- There is a conservative voltage guardband below the nominal level.
- BRAMs power is significantly reduced through Undervolting; however, reliability degrades below min safe voltage.
- We characterized the behavior of Undervolting faults at the critical region.

Future Works

- Dynamic Vmin scaling, adapted by frequency and temperature.
- More advanced designs, where other components such as I/O, DDR, DSP are undervolted.
- Efficient Fault Mitigation Techniques.
- Profiling applications such as Deep Neural Networks (DNNs), among others.
- Extending Undervolting for other commercial FPGAs such as Intel/Altera.

Thanks!

www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Contact:
Behzad Salami
behzad.salami@bsc.es