Understanding Power Consumption and Reliability of High-Bandwidth Memory with Voltage Underscaling

Saber Nabavi **Behzad Salami** **Osman Unsal**

Adrian Cristal **Hamid Sarbazi-Azad** **Onur Mutlu**

Executive Summary

- **DRAM** has problems and one of them is **Bandwidth**.
  - ✓ **HBM** puts DRAM chips inside a package with GPU, FPGA, etc.

- **HBM** uses the package's **power** budget
  - ✓ **Undervolting** reduces power **WITHOUT** losing bandwidth.

- **Push** undervolting **too far**, it will result **unwanted bit-flips**
  - ✓ **This Work: power, bit flips** and **trade-offs**

- **Evaluation Setup**
  - ✓ Xilinx FPGA with **2 Stacks**
  - ✓ **HBM voltage rail**

- **Main Results**
  - ✓ **19% voltage guardband**
  - ✓ **2.3X power savings**
  - ✓ **Fault-map to aid users to take advantage of undervolting**

Outline

- Why HBM?
- What is HBM?
- Undervolting
- Methodology
- Results
  - Power
  - Reliability
  - The Trade-off

Outline

- Why HBM?
- What is HBM?
- Undervolting
- Methodology
- Results
  - Power
  - Reliability
  - The Trade-off

Why HBM?

- **DRAM Limitations:** Power, Latency, Bandwidth, etc.
  - ✓ Specially in **data-intensive** applications
- Replace DRAM: PCM, MRAM, etc.
  - ✓ Have their own **limitations**
- **Improve** DRAM:
  - ✓ Reduced Latency DRAM (RLDRAM)
  - ✓ Graphics DDR (GDDR)
  - ✓ Low-Power DDR (LPDDR)
  - ✓ **High Bandwidth Memory (HBM) -> Bandwidth**
- HBM Use cases
  - ✓ NVIDIA A100, Xilinx Virtex Ultrascale+ HBM, AMD Radeon Pro
  - ✓ The Summit Supercomputer

Outline

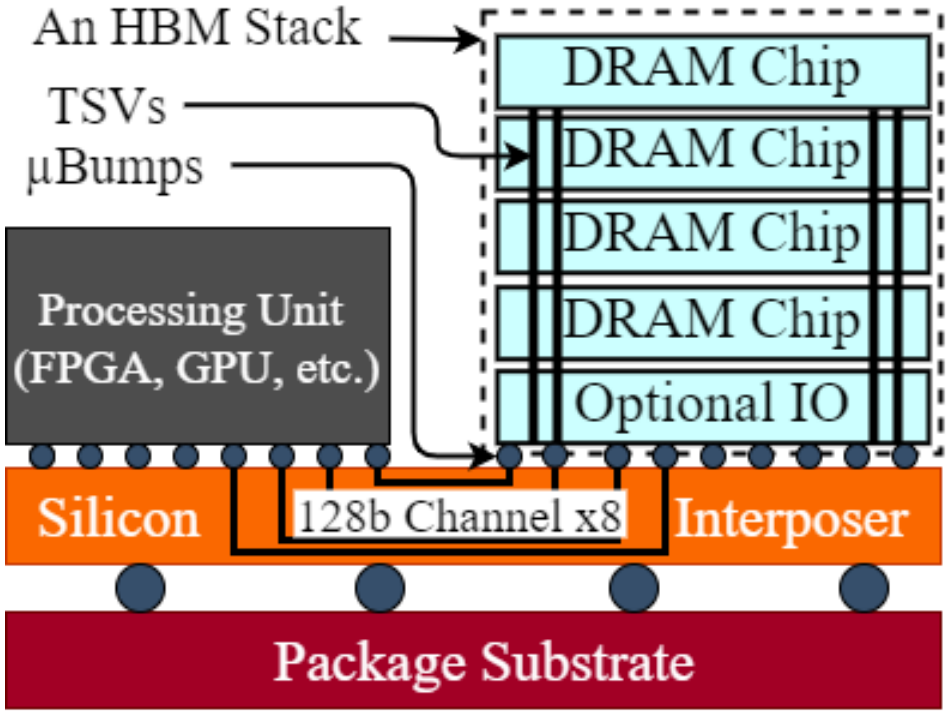
- Why HBM?
- What is HBM?
- Undervolting
- Methodology
- Results
  - Power
  - Reliability
  - The Trade-off

What is HBM?

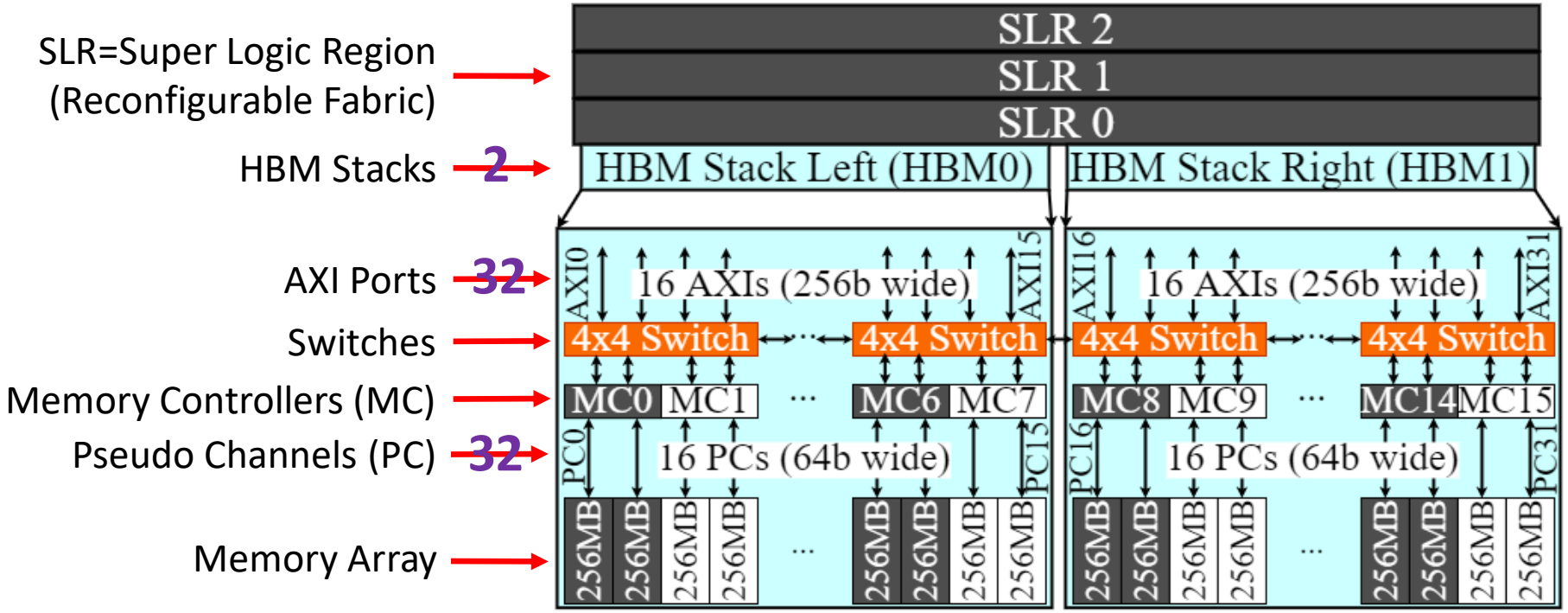
- **IDEA**: Integrate **stacks** DRAM chips into the computing package
  - ✓ Use **TSVs**, **µBumps** and **Silicon Interposer** to connect everything
  - ✓ **Eight 128bits** wide channels per stack

- Benefits:
  - ✓ An order of magnitude **Higher Bandwidth**
  - ✓ **Smaller** form factor
  - ✓ **Lower energy per bit** (7pJ vs 25pJ in DDRx)

- Challenge:
  - ✓ Uses the package's **power** budget
  - ✓ Save power but **NOT** lose bandwidth: **Undervolting**



Xilinx VCU128



Outline

- Why HBM?
- What is HBM?
- Undervolting
- Methodology
- Results
  - Power
  - Reliability
  - The Trade-off

Undervolting

- **IDEA: Reduce supply voltage but keep Frequency**
  - ✓ We can do this because of **Voltage Guardband**
  - ✓ **Save power WITHOUT losing bandwidth**

- **Catch:**
  - ✓ **Pushed beyond guardband, bit flips will appear!**
  - ✓ **But we can save even more power at the cost of these faults!**

- **Our Work:**
  - ✓ **Undervolt HBM**
  - ✓ **Then push in too far!**
  - ✓ **Report power saving and bit flips**
  - ✓ **Trade-off among memory capacity, power and fault-rate**

Outline

- Why HBM?
- What is HBM?
- Undervolting
- Methodology
- Results
  - Power
  - Reliability
  - The Trade-off

Methodology

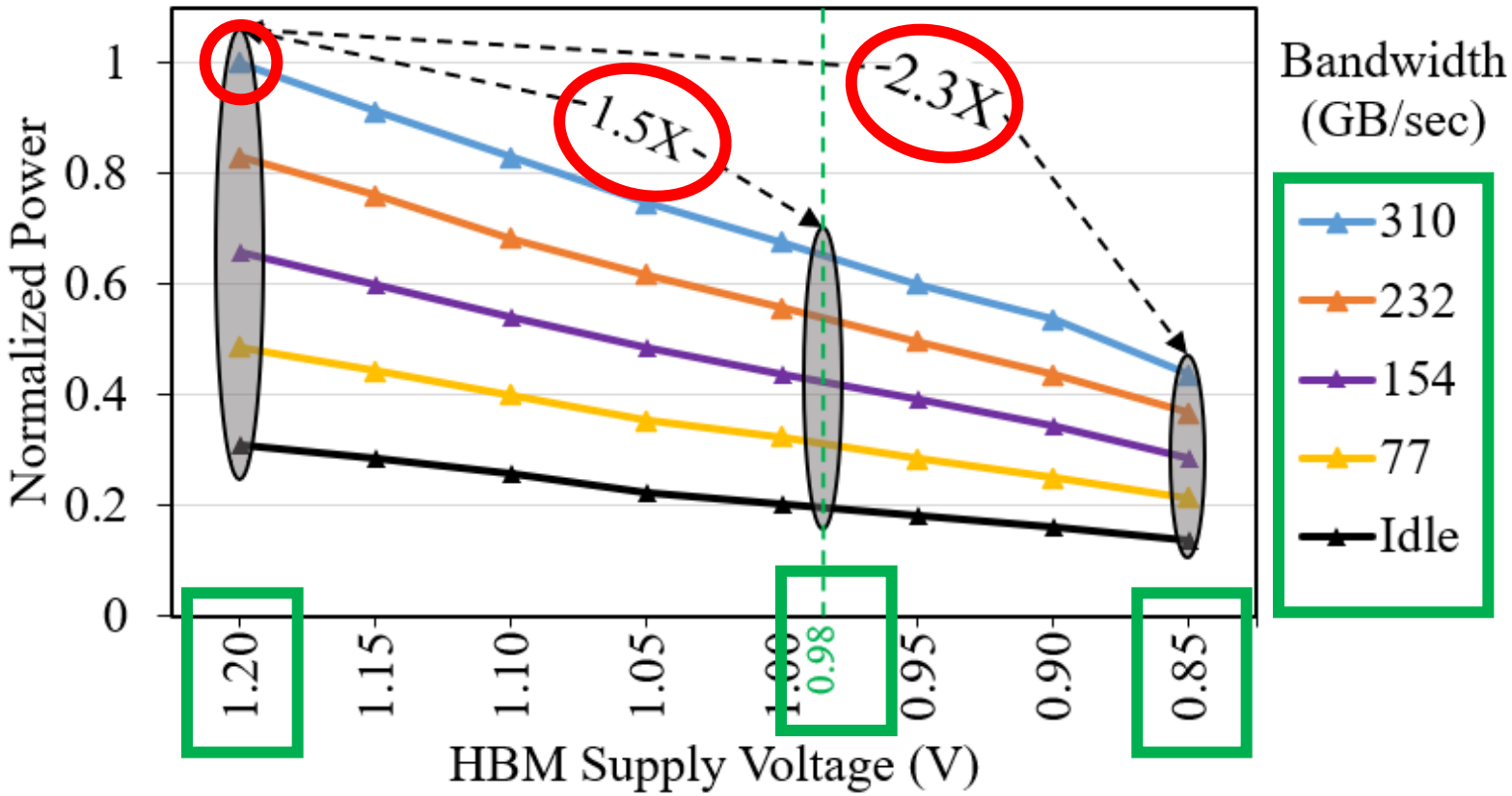
- Undervolting Mechanism
  - ✓ HBM supply voltage is driven by an on-board regulator
  - ✓ We control it from the host
  - ✓ 10mV voltage steps
- Power Measurements:
  - ✓ Change bandwidth utilization by enabling/disabling AXI ports
  - ✓ Measure power at all voltage steps
  - ✓ Measure idle power by disabling all AXI ports
- Reliability Test
  - ✓ Test the entire memory vs. pseudo-channel
  - ✓ **Write all 1s** (to detect **1-to-0** bit flips)
  - ✓ **Write all 0s** (to detect **0-to-1** bit flips)

Outline

- Why HBM?
- What is HBM?
- Undervolting
- Methodology
- Results
  - Power
  - Reliability
  - The Trade-off

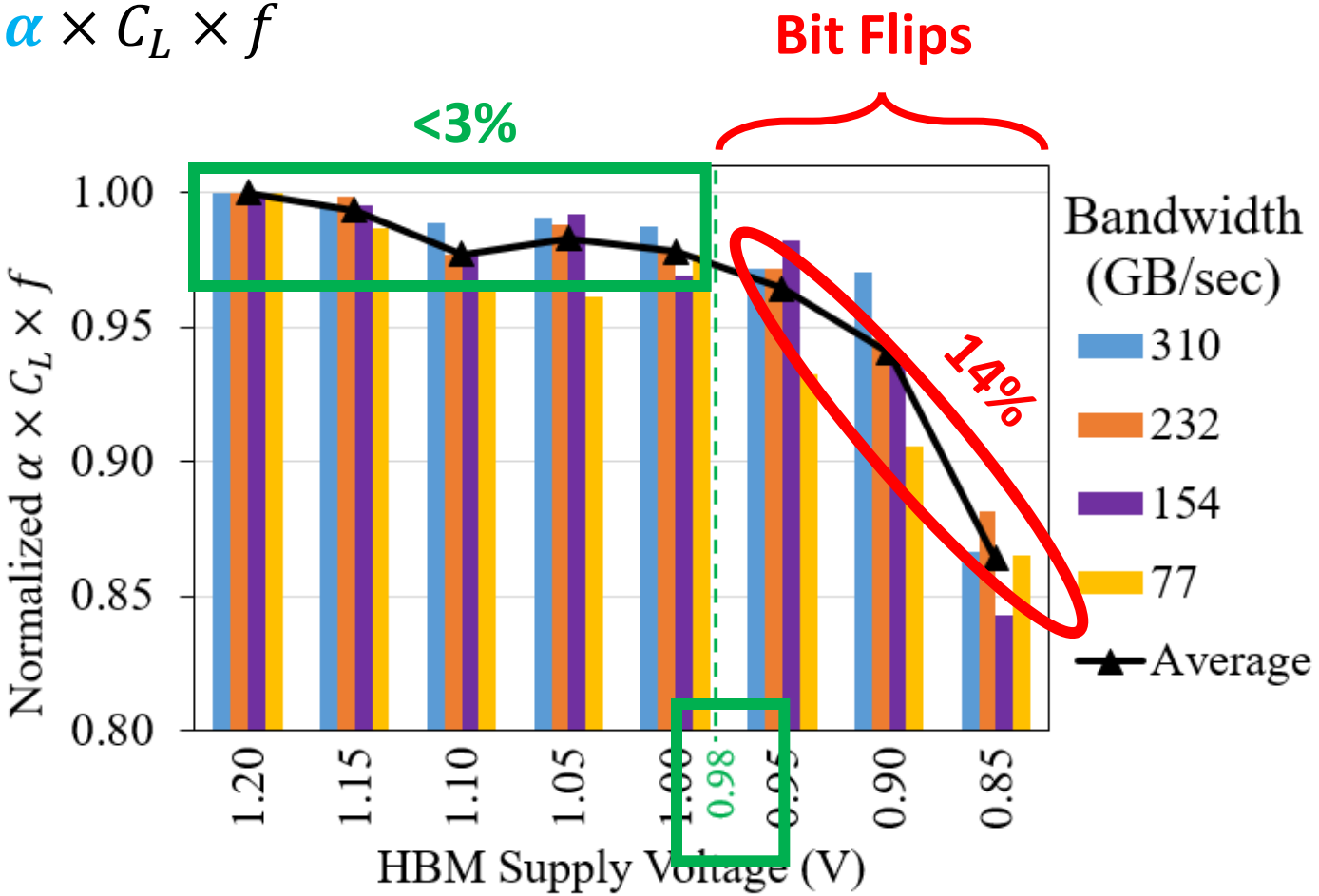
Power: Active and Idle

- Active Power $= \alpha \times C_L \times f \times {V_{dd}}^2$



Power: α

- **Active Load Capacitance** $= \alpha \times C_L \times f$
- Unit: ***farads per second***
- f and C_L are constant

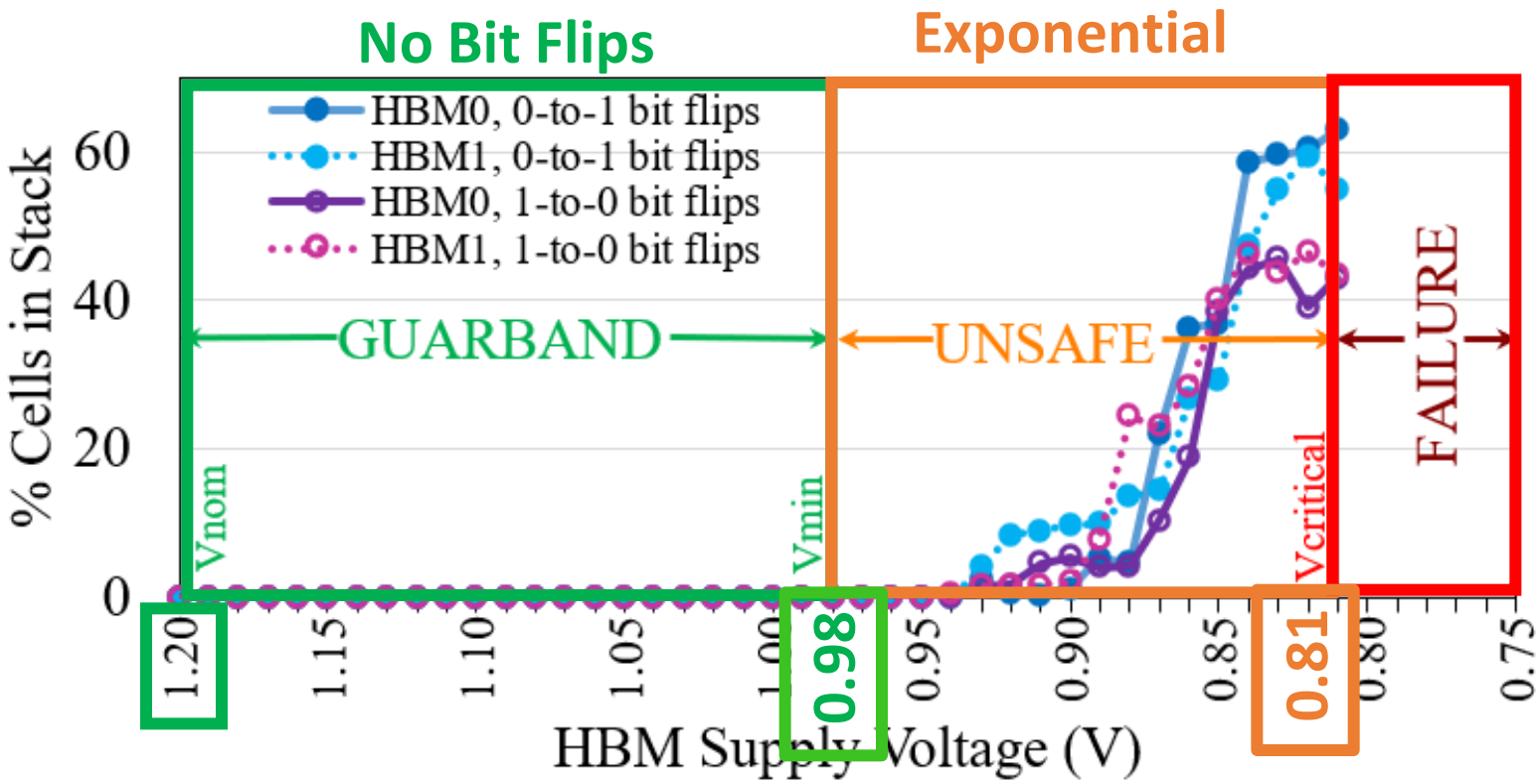


Outline

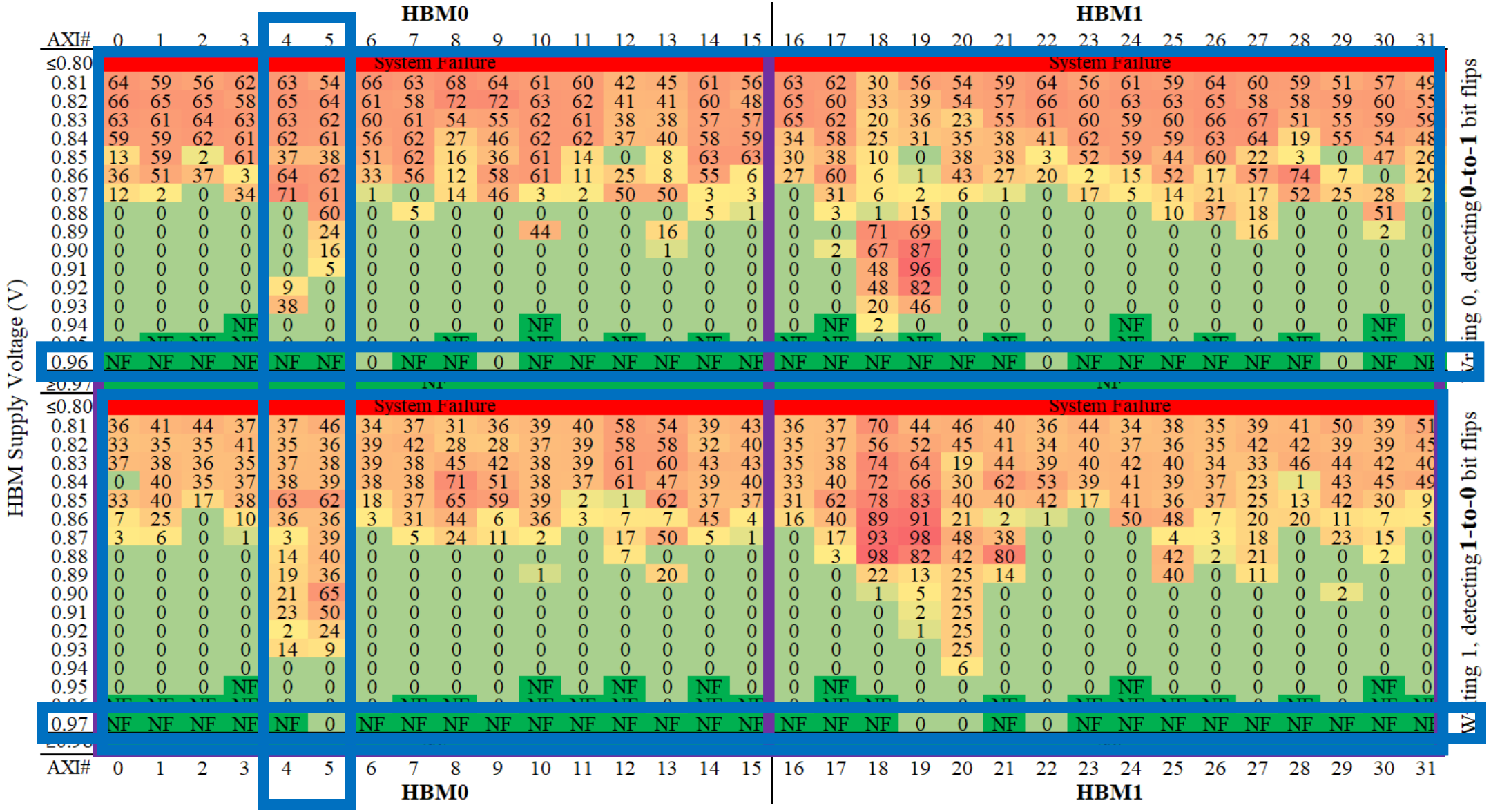
- Why HBM?
- What is HBM?
- Undervolting
- Methodology
- Results
  - Power
  - Reliability
  - The Trade-off

Reliability: the Regions

- **Guardband**:
  - ✓ Below nominal level ($V_{nom} = 1.2V$)
  - ✓ No faults
- **Unsafe**:
  - ✓ Below guardband ($V_{min} = 0.98V$)
  - ✓ **Exponential** growth in bit-flips
  - ✓ Max out at **0.84V**
- **Failure**:
  - ✓ HBM **crash** point ($V_{critical} = 0.81V$)
  - ✓ Restart with nominal voltage



Reliability: the Variations



Outline

- Why HBM?
- What is HBM?
- Undervolting
- Methodology
- Results
  - Power
  - Reliability
  - The Trade-off

The Trade-Off

Understanding Power Consumption and Reliability of High-Bandwidth Memory with Voltage Underscaling

Saber Nabavi **Behzad Salami** **Osman Unsal**

Adrian Cristal **Hamid Sarbazi-Azad** **Onur Mutlu**