Background:
- Near Threshold Computing (NTC) [1] improves energy efficiency by reducing voltage.
- Performance impact from NTC can be mitigated through parallelism.
- GPUs are ideal for such scenario.
- NTC requires more current, thus making the system more sensitive to process variation (PV) [2].
- Increased current also reduces power delivery efficiency.

Voltage Stacking:
- Instead of the parallel power delivery, use a serial approach.
- In a N-stack level, supply voltage is multiplied by N, and current drops by around N.
- Challenge: keep the load balance of the stack levels.

Voltage Stacking for Process Variation Compensation:
- Channel length (Left) or an threshold voltage (Vth) increase will result in higher impedance and slow device.
- Conventional: PV results in a lowered current through the core. This results in a higher delay and a slower core. To compensate, higher voltage can be applied.
- Stacked: the same current passes through the stack, therefore, higher impedance results in a higher voltage across that core, which naturally compensates the PV effects.

Case study: Inverter chain

Voltage Stacking reduces the overall system current, and thus reduces the pressure on the power delivery.

What to stack?
- Stacking needs balanced load ➔ Stack equal structure.
- Stack GPU PEs (lanes) within a SM.
- Within a SM, lanes work in lock-step, running the same program.
- Foot/Head position is fixed during design time for simplicity.
- Compensation works well for cores with reverse variation.
- Stack most positive with most negative variation.
- Cluster multiple lanes per SNET.

Evaluation Setup:
1) Expected compensation over 10k GPU dies:
- Varius-NTC [3] to generate 10k variation maps.
- Calculate expected voltage (from impedances).
- Varius-NTC [3] to calculate the new delay/power.
2) Check power delivery quality:
- Generate a time varying impedance model for each core.
- IBM PowerGrid Benchmarks to model power grid.
- SPICE simulation of the power grid with the core models.

Overal Results:
- GPU Stacking did not hurt power delivery quality.
- Stacking delivers 80% of the performance compared to the no variation conditions.
- GPU Stacking did not hurt power delivery quality.
- GPU Stacking reduces IR drop.
- Reduces the pressure in power grid design.
- GPU Stacking is also able to compensate PV effects.

Conclusions:
- GPU Stacking manages PV.
- Stacking can increase performance under PV at NTC on average by 37%.
- Stacking delivers 80% of the performance compared to the no variation conditions.
- GPU Stacking did not hurt power delivery quality.
- GPU Stacking reduces IR drop.
- Reduces the pressure in power grid design.
- GPU Stacking is also able to compensate PV effects.