SWTEST

PROBE TODAY, FOR TOMORROW

2024 CONFERENCE

Challenges of Heat Generation in Probe Cards

Micron®

Intelligence Accelerated™

Kurt Guthzeit

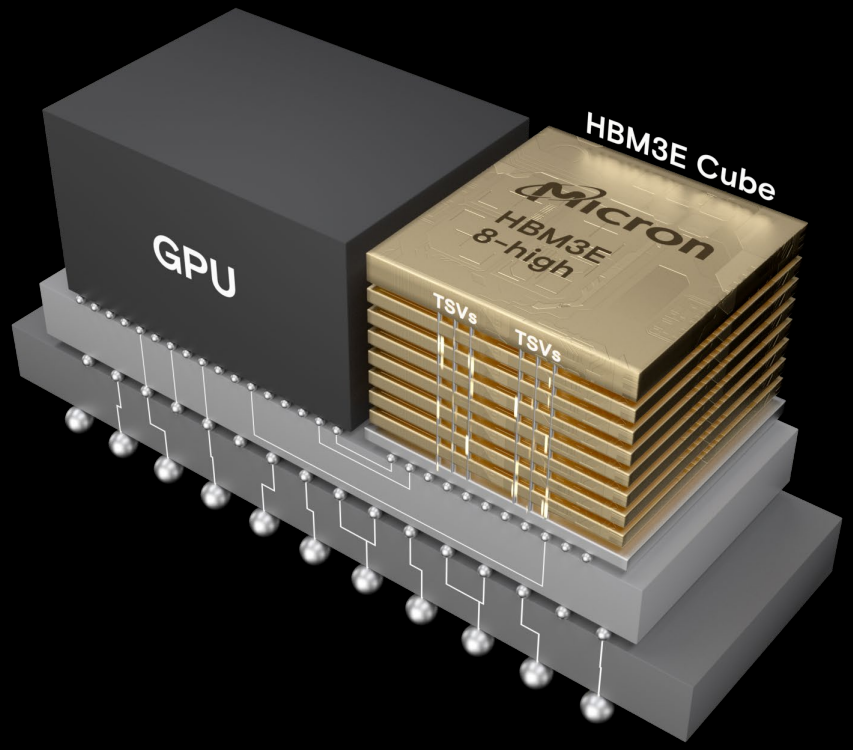
Micron Technology

Overview

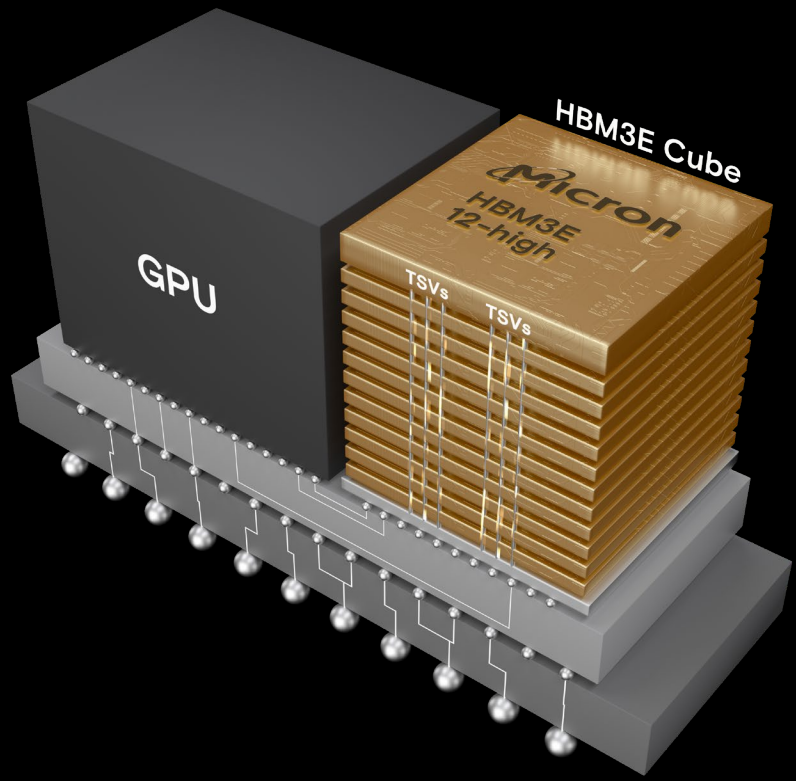
- HBM landscape
- HBM: High Performance Memory use case
- Probing challenges of Advanced Packaging
- Heat generation in probe cards
- Conclusions

Micron HBM3E

Industry-leading performance, market-leading lower power consumption



8-high, 24GB in Volume Production



12-high, 36GB Now Sampling

24GB & 36GB

enables higher precision training

>1.2TB/s

fuels AI compute cores

~30%

Lower power consumption than competition

Demands of AI Infrastructure

Performance

Large Language Models (LLM) which are the foundation for generative AI applications achieve <50% of peak processor performance due to memory bottlenecks[1]

Capacity

LLM sizes and datasets continue to grow exponentially requiring multiple servers to fit the model parameters and tokens

Power

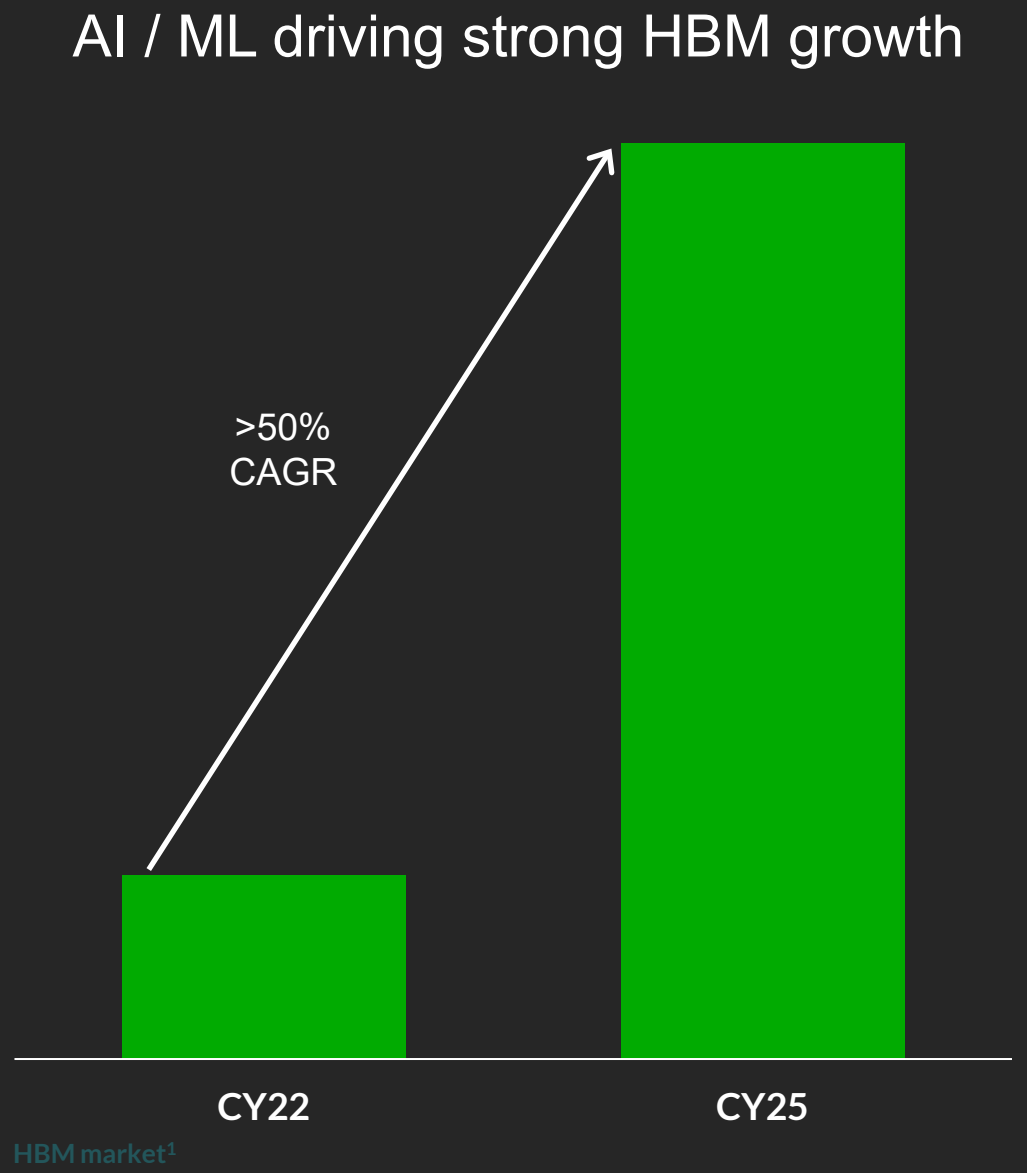
Data centers consume ~3% of the world's electricity, and are expected to reach >8% by 2030[2]

Training LLMs such as GPT-3 can consume upwards of 1GWh energy[3]

1, https://arxiv.org/abs/2104.04473 2. https://www.eetimes.eu/power-management-facilitating-the-energy-journey/ 3. 2104.10350.pdf (arxiv.org)

Explosive growth demands innovation

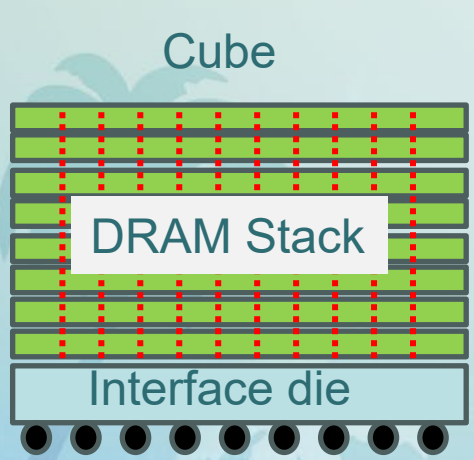
- Productivity enhancing applications driving AI hardware boom, fueled by HBM
- AI focused servers driving overall memory and storage content growth vs. general-purpose servers
  - 6x-8x DRAM Content

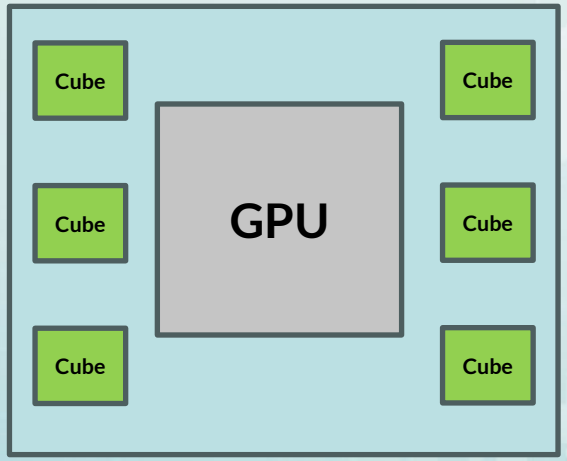


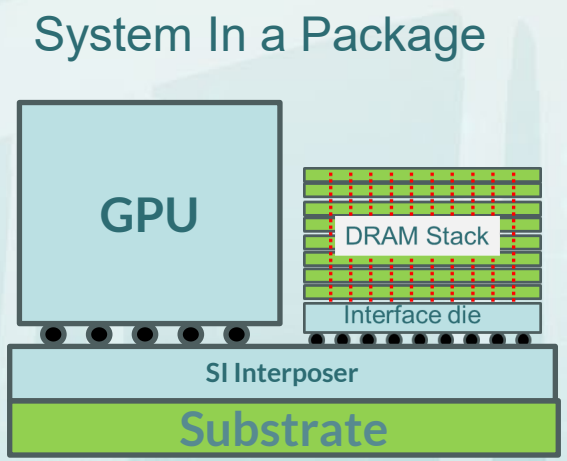
1 OMDIA CY22, Micron Market Model CY25, All HBM

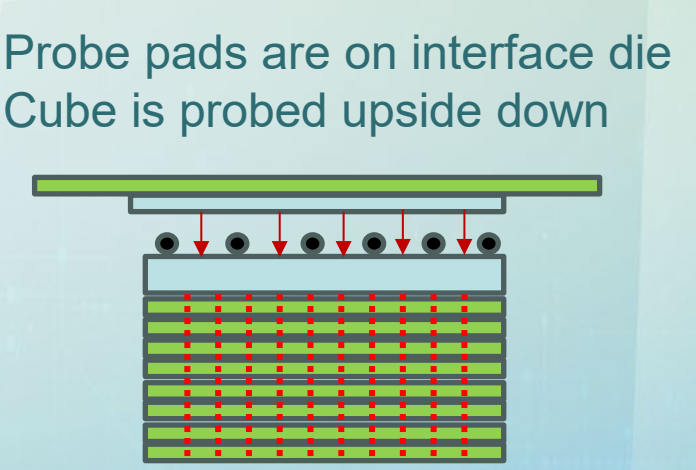
HBM: High Performance Memory

- High-bandwidth memory (HBM) utilizes stacking of DRAM die for expansive bandwidth and low power per transaction.
- HBM3 has 1024 I/O's and an interface die collecting data from DRAM layers to send out on multi lane highway.
- Conventional DRAM has 16 to 128 I/O's can only send data on limited highway.
- HBM3 1.2TB/s ~10x bandwidth of GDDR6 128GB/s.





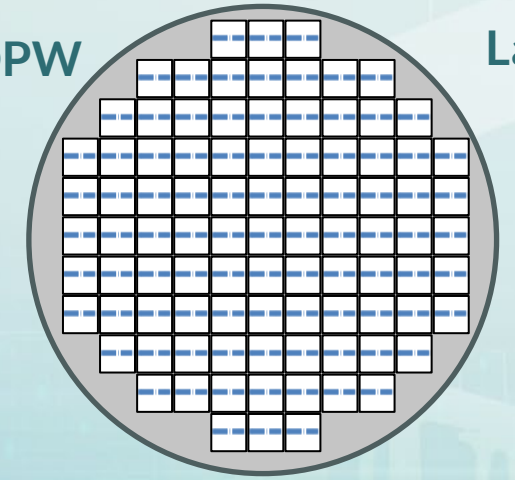




Contrasting Memory Use Cases

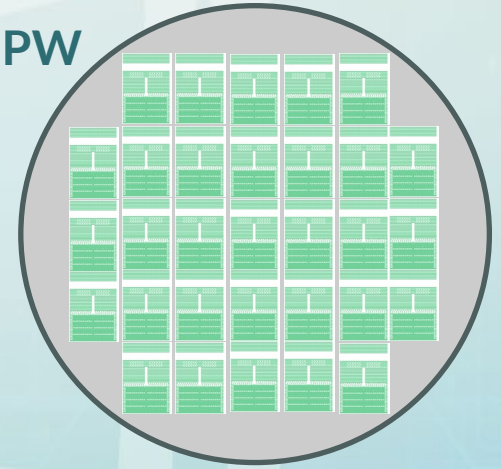
Characteristic	Use Case 1	Use Case 2	OK?
Duts per wafer	1-2K	>6K	✓
Pitch	Driven by die size	JEDEC + test strategy	✓
Pad size	Driven by die size	JEDEC + test strategy	✓
Probe count	Driven by test modes	JEDEC + test strategy	✓
Pad layout	SLOC	Grid Array + bumps	!
Thermal control	Understood	Evolving (power)	!
Test content	Traditional	Evolving (interconnect)	!

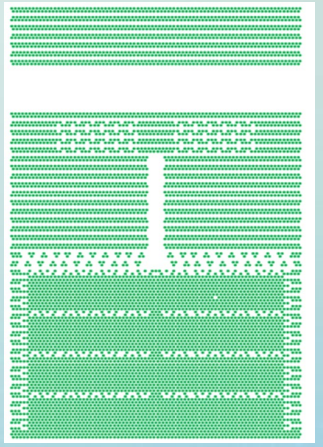
Small Die – high DPW



SLOC – Bond Pad

Large Die – low DPW





Grid Array + Bumps

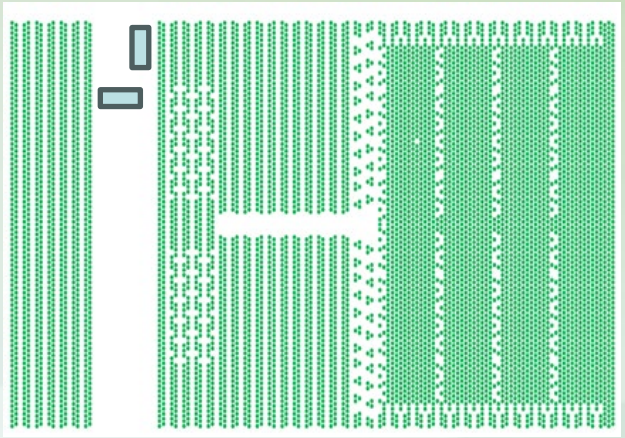
Advanced Packaging: 1TD Probing Challenges

- **System considerations**
  - Chuck force requirements
  - Optimal probe placement for test content
  - Probe card temperature control with high currents

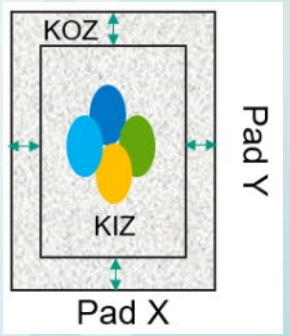
- **Evolving test content**
  - Approximation of end user application conditions
  - Increased / compressed test content
  - TSV, functional, reliability, speed, stress
  - Multiple test step insertions

Advanced Packaging: Probe Layout Challenges

- **The target is shortest test time**
  - Maximized power = heat
- JEDEC defines the bump matrix
- Layout drives four-sided probing
- Vert & Horiz pads compromise SBP
- Complexity of avoiding uBump collision
- **Probe card scaling challenges**
  - Not a solid silicon wafer



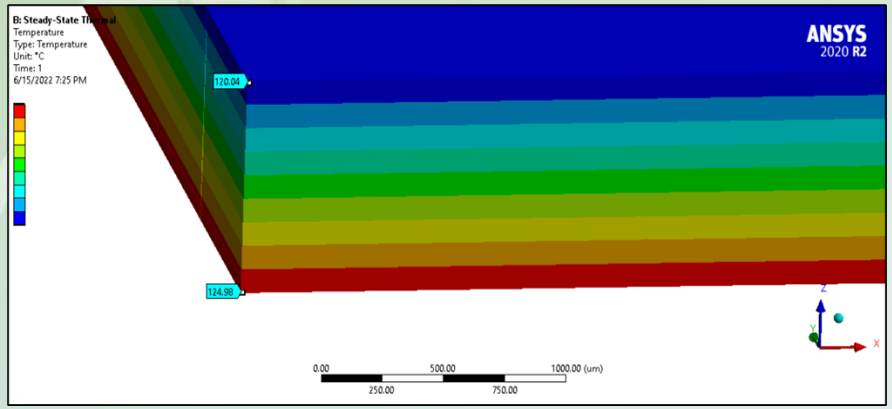
Super Bond Pad



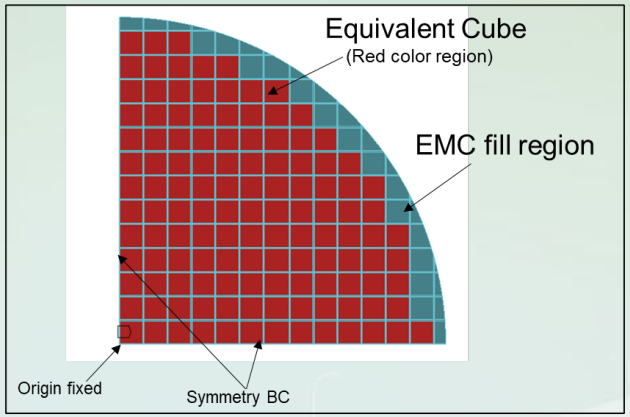
Bump Strike

Modeling for Scaling

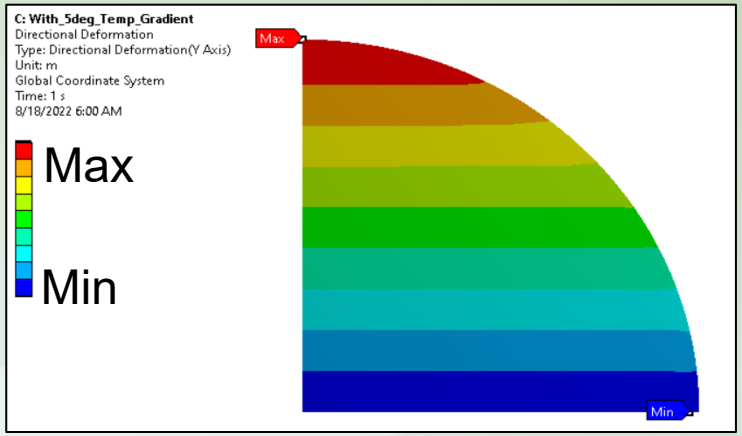
Temp Gradient through Stack



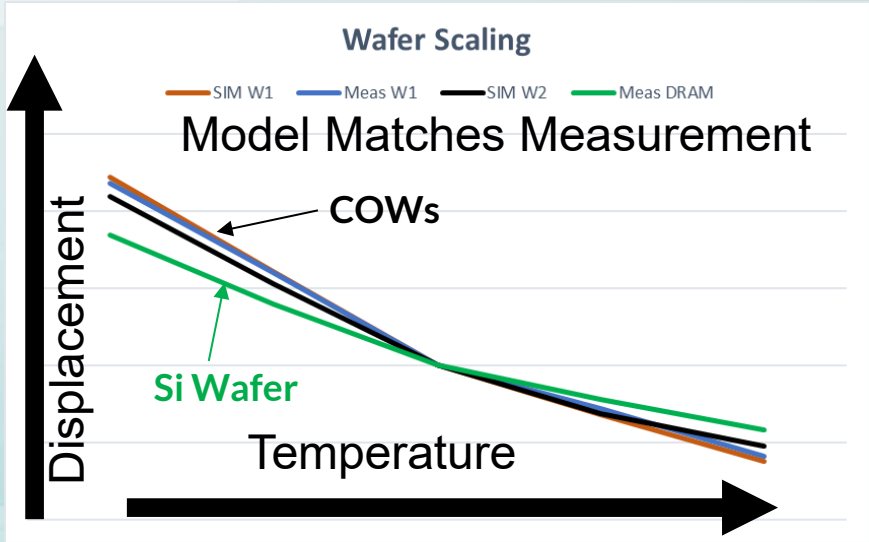
¼ Model



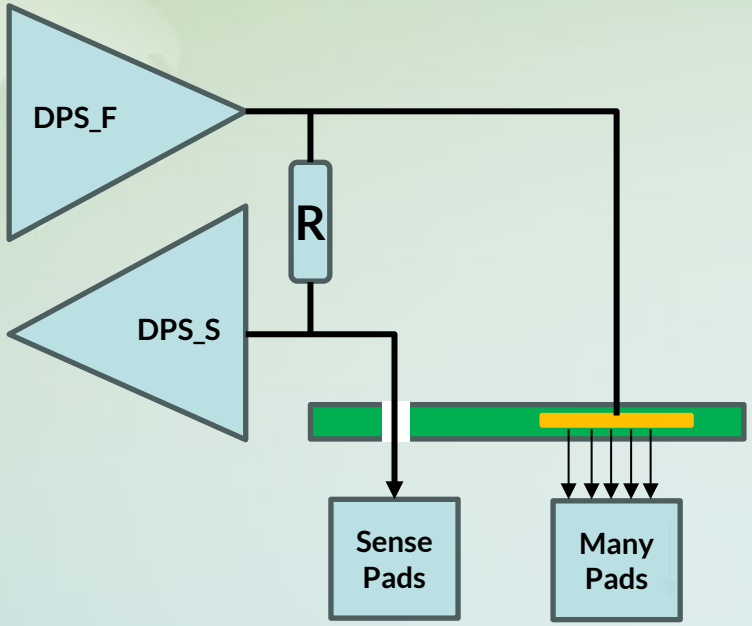
Wafer directional displacement at HT

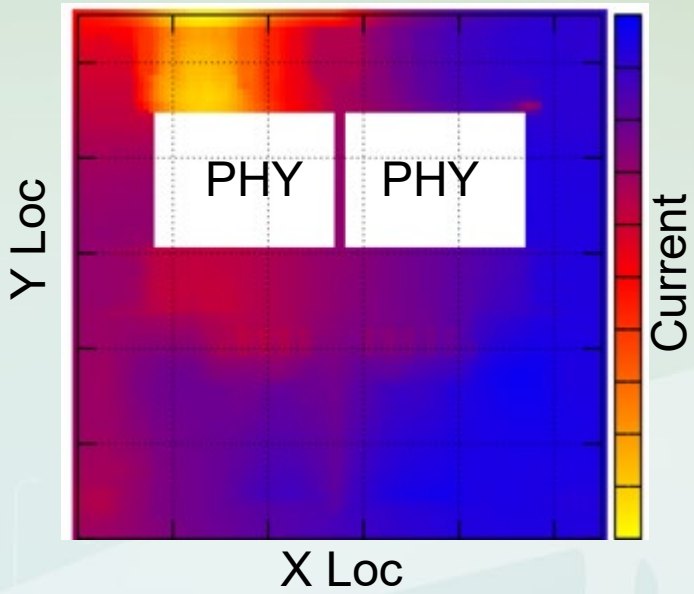


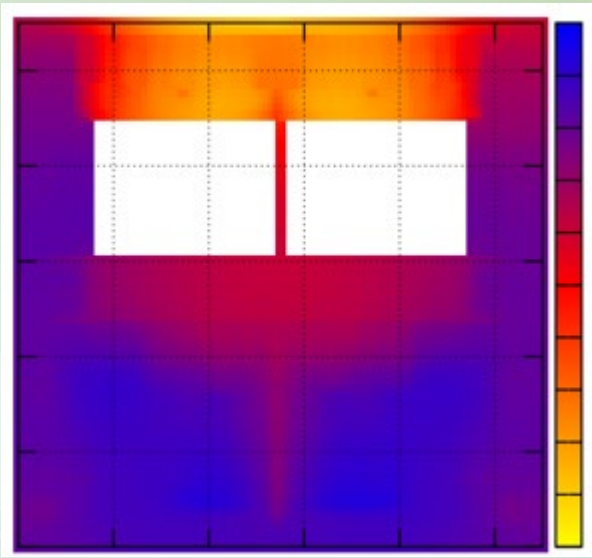
- Previous product expansion was measured at 5 temps to capture non-linearity.
- ¼ model built and expansion validated against measured values.
- Model updated with new material properties and used for probe card scaling.



Optimizing PDN for High Quality Test

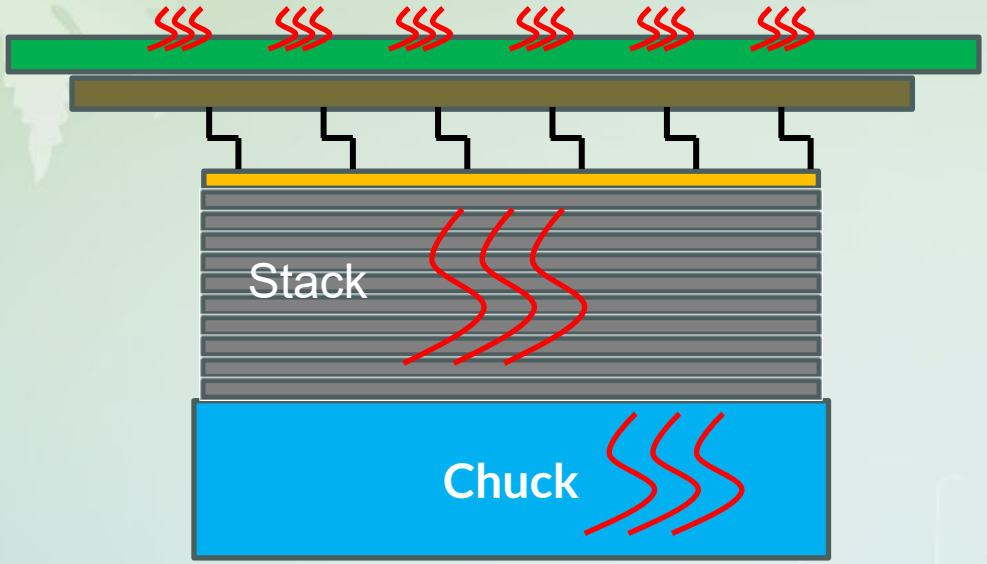


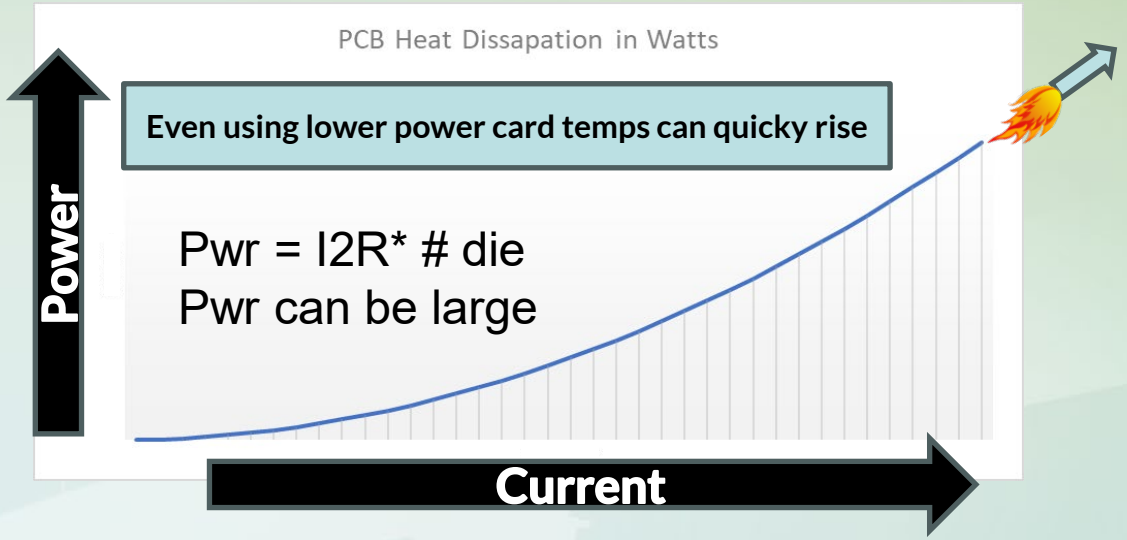




- Spatial power SIMs determine best PWR/GND pad and probe locations.
- Hot spots need to be considered for all test patterns.

Heat Sources

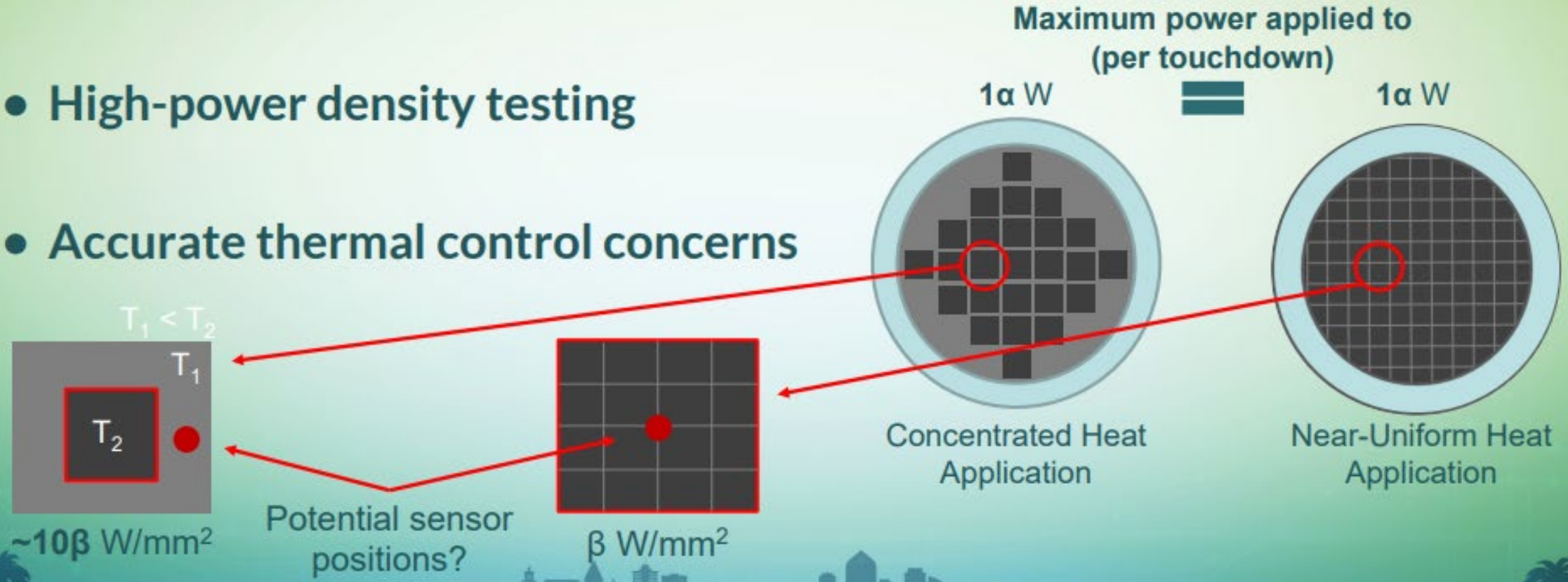




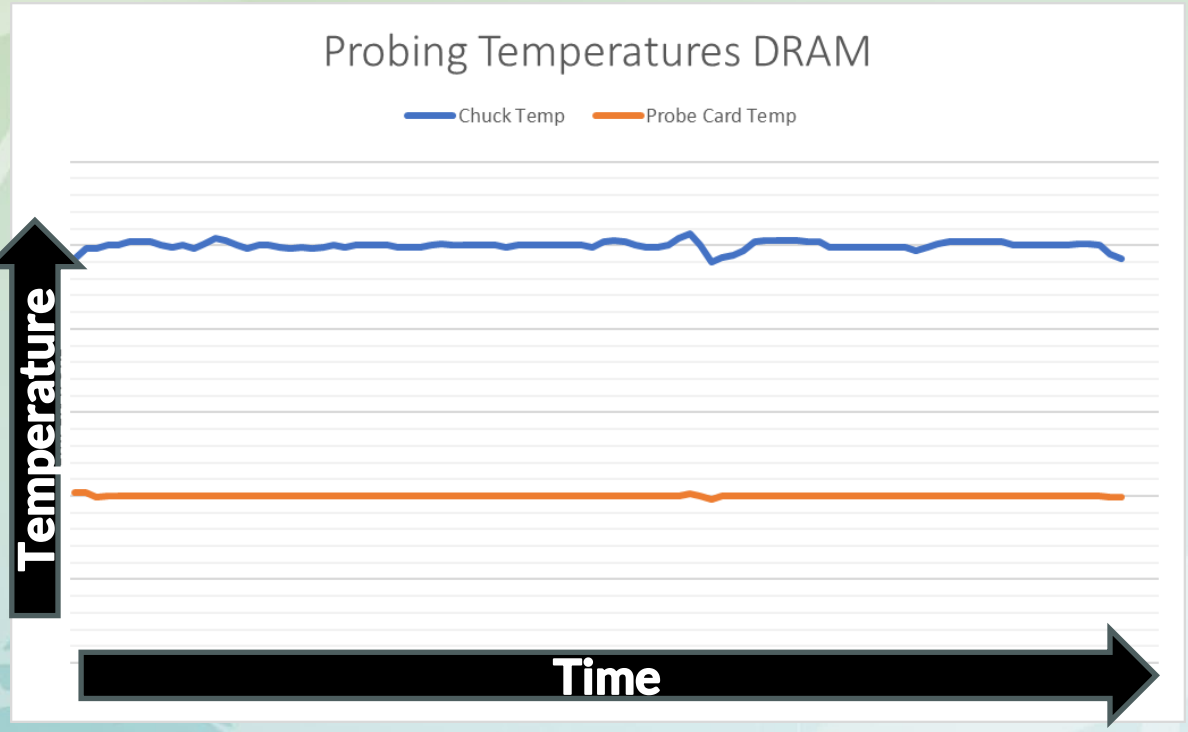
- Heat sources in probe are normally the chuck and the wafer.
- The probe card in certain conditions can be an additional heat source.
- Card heating will affect SBP and potentially damage the card.
- Power is controlled to ensure optimal quality test.

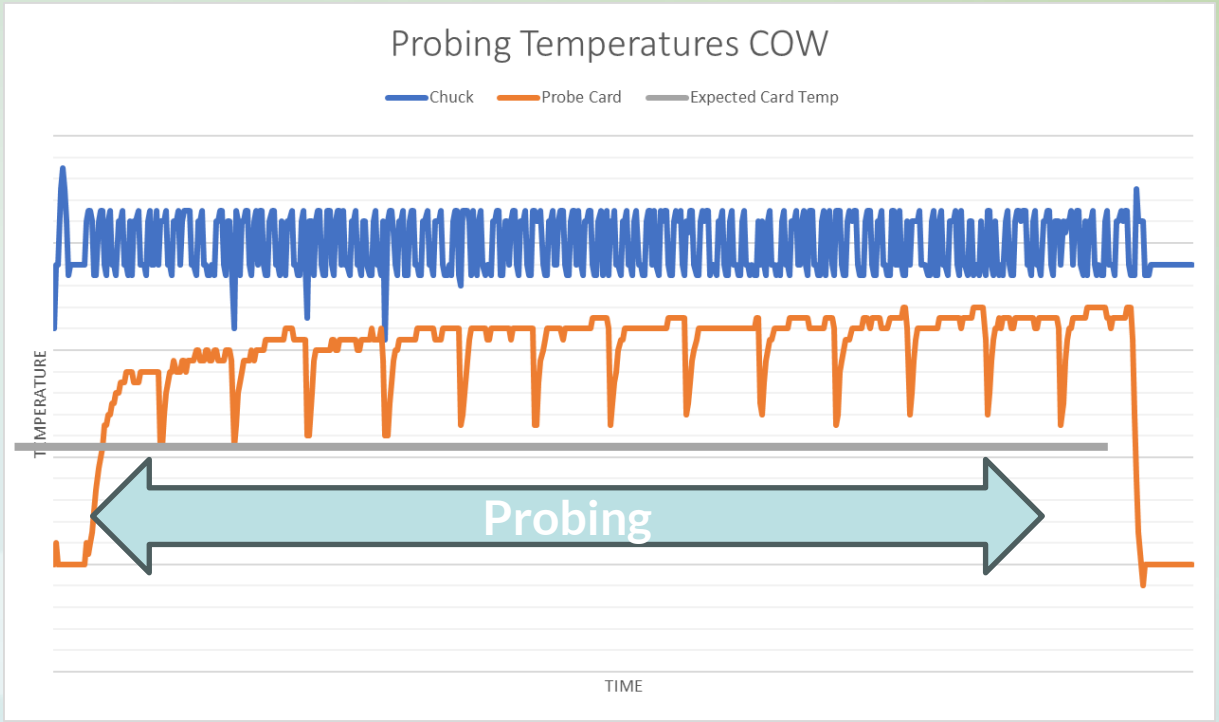
What makes memory different?

- Explosion in memory testing conditions & requirements
- High-power density testing
- Accurate thermal control concerns



Probe Card Ohmic Heating

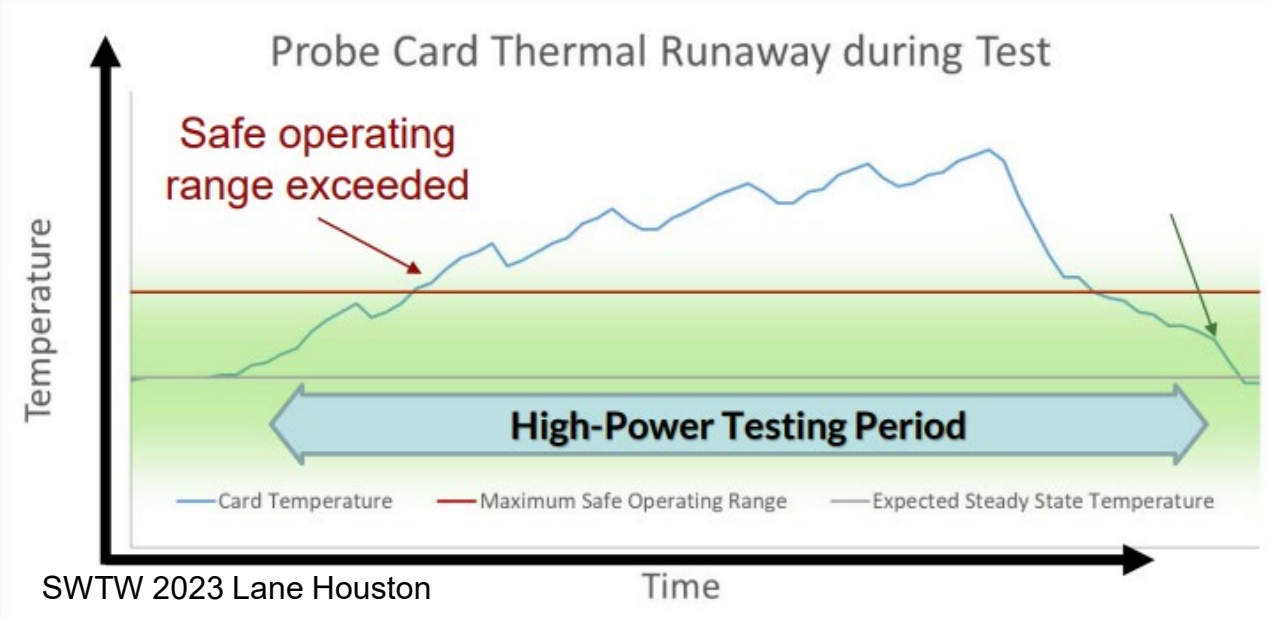


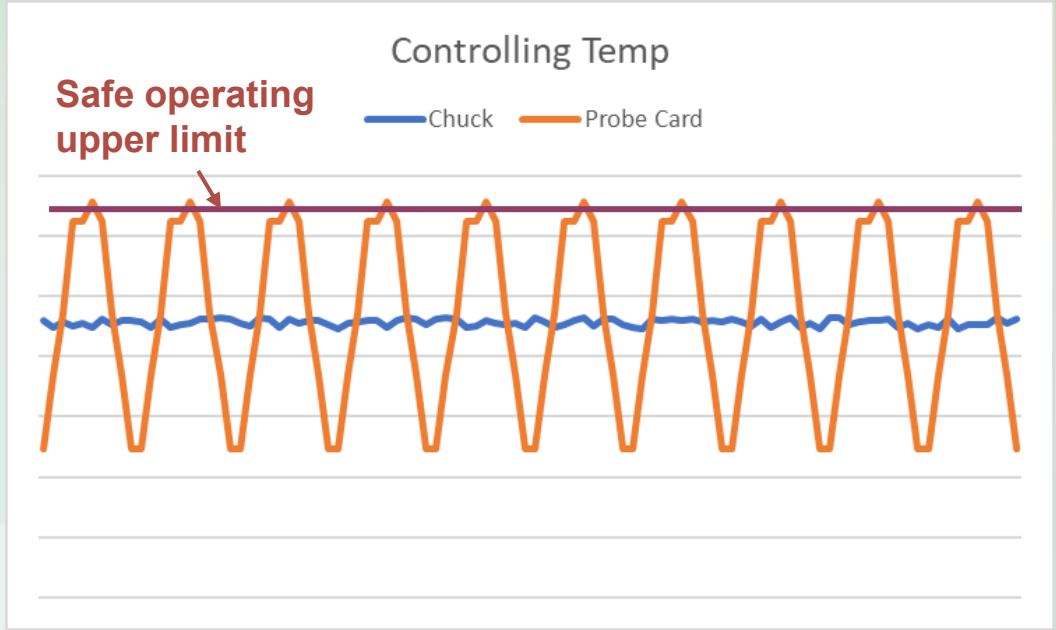


- Traditional Wafer
- Chuck & probe card temp stable

- Probe card ohmic heating at high power test
- Within limits but constrained

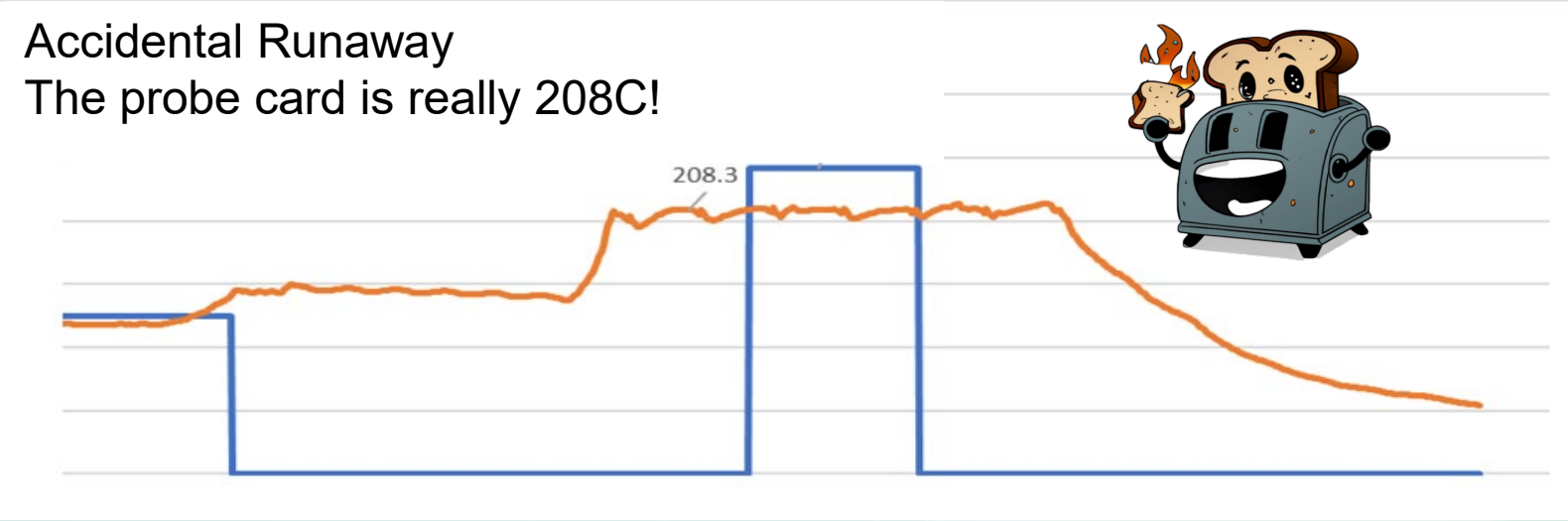
Ohmic Heating Exceeding Spec





- In some test cases card temp can exceed safe operating range.
- Constrained by current probe card technology.
- Controls in place to protect probe card cost test time.

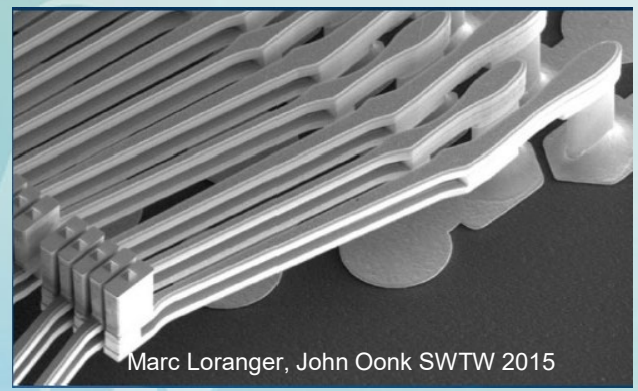
Monitoring Test Cell Health

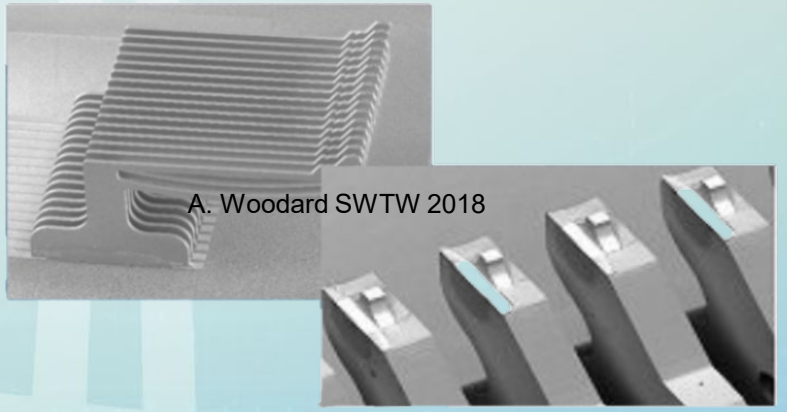


- Triggering action for OOC events
  - Chuck temp: Control with pause or multi-pass testing
  - Chiller health: Disallow probing
  - Probe card critical over temp: Shut down
  - Test program monitors card/chuck temp and act

Limitations of Industry Offerings

- Probe Card Basics Matter More than Ever
  - Probe Tip Cleaning: Production test efficiency
  - Probe Tip Condition: Low Cres, test @ speed
  - Probe Robustness: Tolerant to foreign environment
  - Probe Layout: Complex probe placement
  - Creep at extended time at temp and high current





Conclusion

- Probing advanced memory highlights new challenges
  - Opportunity for further industry innovation & collaboration
- Ohmic heating is a new issue for memory probe
  - Evolving test content + stacked wafers consumes more power
- Thermal monitoring and reaction is important
  - Wafer side temperature is more critical
- Call to Action: Partner with Memory Maker
  - No single probe solves all issues

Thank You!

Acknowledgments:
Micron: Aaron Woodard, Alistair Laing, Emmett Ricks, Lakshmi Narayana Kurra, Thiagarajan Raman