Embedded Microfluidic Cooling for Nextgen High-Power Server Architectures



Michael Cumbie
Distinguished Technologist
Primary Investigator
michael.cumbie@hp.com



Paul Benning
HP Senior Fellow
Tech 2 Market Lead
paul.benning@hp.com



Tahir Cader
NVIDIA Principal Thermal Architect
tcader@nvidia.com



Tom Gray
NVIDIA Circuit Research Senior Director
tgray@nvidia.com

COOLERCHIPS™

Project Vision

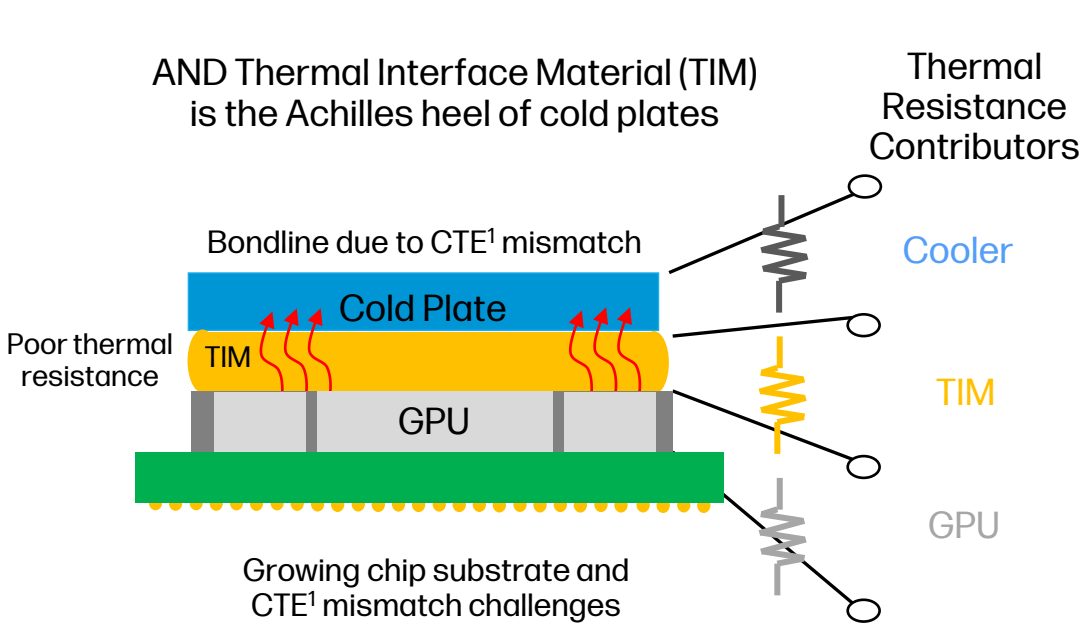
- Our microfluidic Silicon Cold Plate (SiCP) solution will be smaller, lighter, and will deal effectively with die warpage through microbond technology developed at HP: *A microfluidic cooling solution that scales with GPU power and size roadmap*
- This will be achieved by leveraging HP's commercialized 5th generation of inkjet microfluidics to directly couple microfluidics to NVIDIA's GPU chip surface

PROTECTED DATA - HP Confidential, not for public disclosure

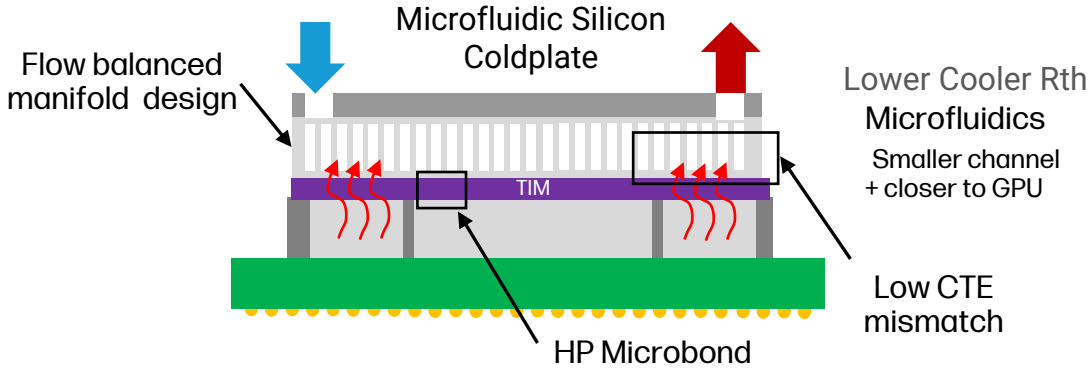
HP silicon-based chip cooling technology approach

Dissipate heat more efficiently and quickly while scaling with the size of the chips

Traditional Coldplate manufacturing is more challenged to take advantage of microfluidics



HP Approach: Silicon microfluidics that scale with GPU power and size requirements



Notes: 1) CTE: Coefficient of Thermal Expansion 2) MEMS: Micro-electromechanical systems

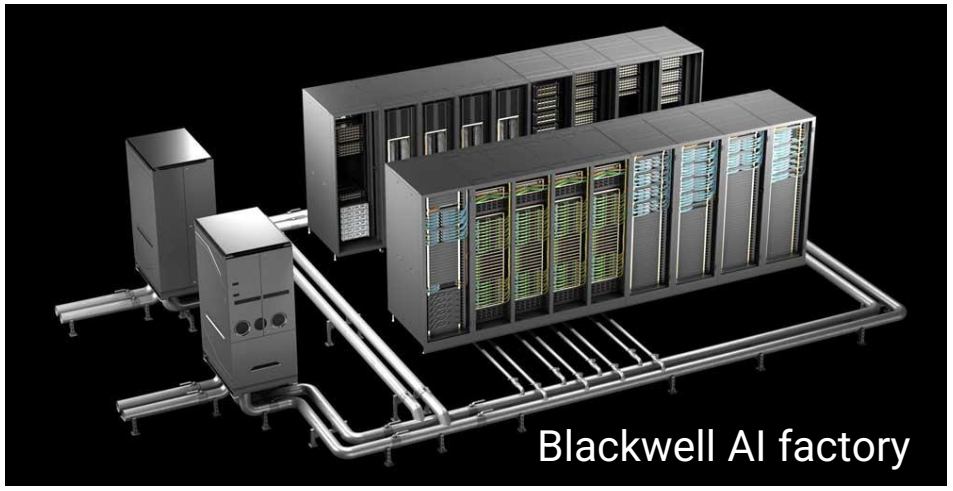
arpa·e CHANGING WHAT'S POSSIBLE

PROTECTED DATA - HP Confidential, not for public disclosure

Solution / Product / Technology Description And Impact

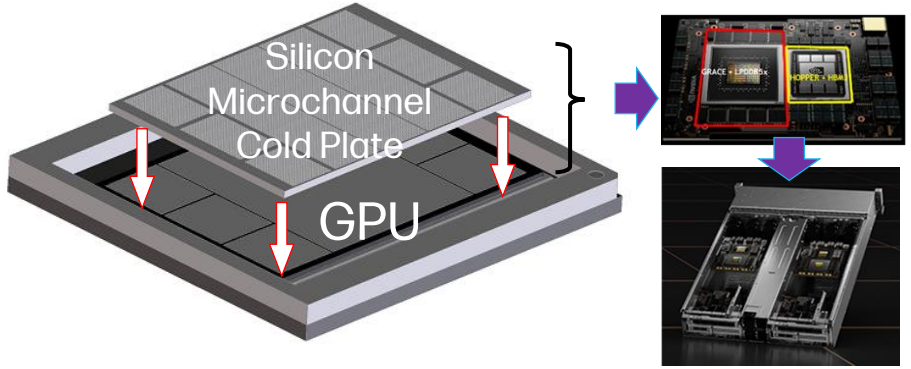
- ***We are developing a single-phase cooler that will reject server heat to 40°C and 60% relative humidity external ambient air***
- **Compared to the state-of-the-art** our microfluidic Silicon Cold Plate (SiCP) solution will be smaller, lighter, and will deal effectively with die warpage through microbond technology developed at HP.
- **Our SiCPs will be designed as a drop in solution** for single-phase, liquid cooled systems, accelerating time to market by leveraging HP's G5 platform, developed over the last decade for high volume commercial printing applications that also demand an extremely high level of reliability.

Drop in solution for single-phase systems



Img src: https://nvidianews.nvidia.com/online-press-kit/computex-2024-news

Silicon Microchannel Cold Plate + Nextgen GPU

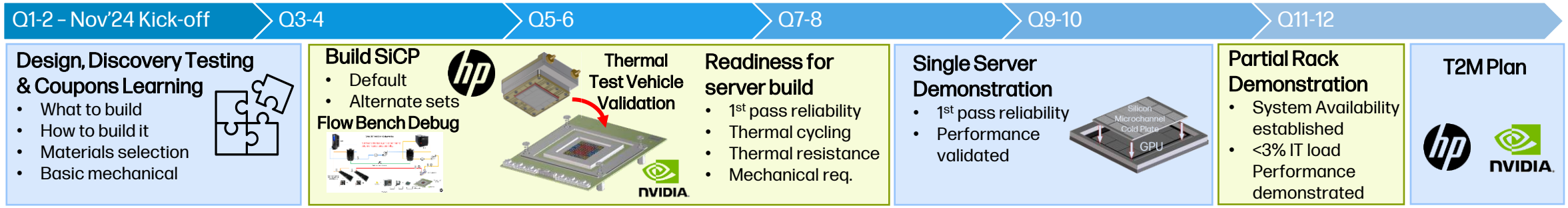


arpa·e
CHANGING WHAT'S POSSIBLE

PROTECTED DATA - HP Confidential, not for public disclosure

Task Outline & Technical Objectives

- ***Technology Objectives***
  - *Develop advanced microfluidic silicon cold plate (SiCP) technology for Nextgen high-power GPUs and server architectures.*
  - *Demonstrate scalability of the SiCP to high volume manufacturing, and that it can meet reliability and cost targets.*
  - *Demonstrate an SiCP integrated into a single, test server and partial rack.*
- ***Development Activities***
  - *HP: Silicon microchannel cold plate design, modeling, manufacturing, reliability, fluidic/thermal performance testing*
  - *NVIDIA: Server and rack integration, performance testing and reliability analysis; IT System modeling*



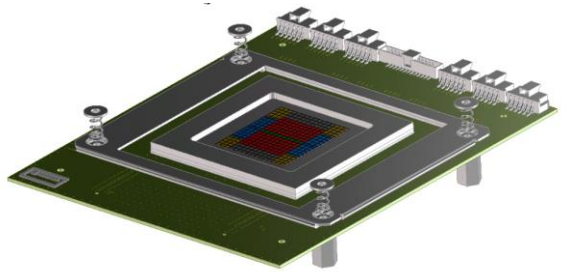
arpa·e CHANGING WHAT'S POSSIBLE

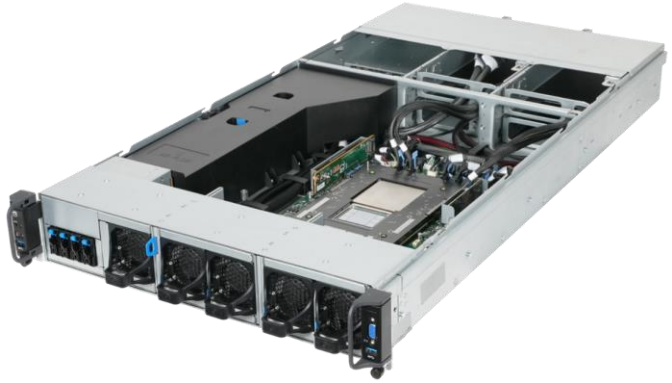
PROTECTED DATA - HP Confidential, not for public disclosure

Target Learning System for Risk Retirement

- TTV
- Hopper Platform

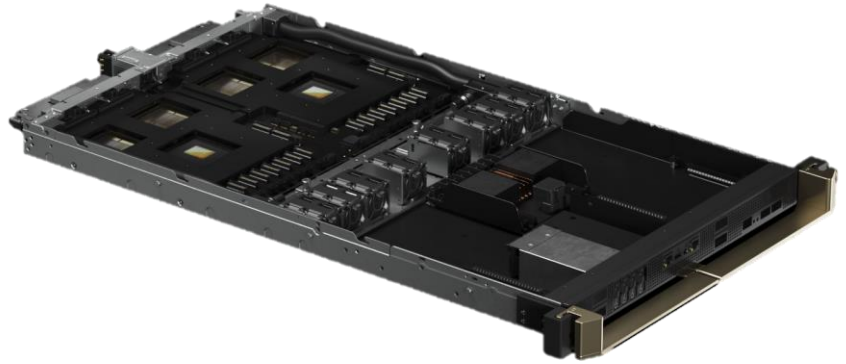
Learning Testbeds to retire risk ahead of target Server Integration





- Blackwell Platform

COOLERCHIPS Proof of concept Server integration



PROTECTED DATA - HP Confidential, not for public disclosure

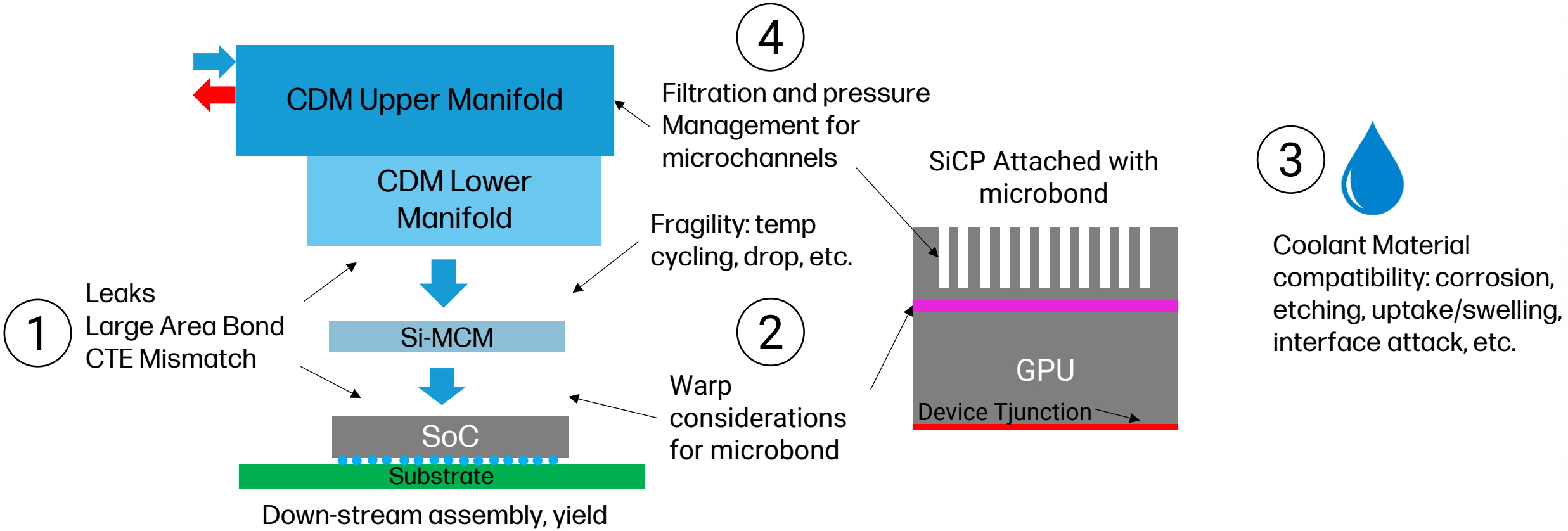
High-level Milestone Deliverables

Nov'24 Kickoff	M1 / Q1	M2 / Q3	M3 / Q4	M4 / Q6	M5 / Q9	M6 / Q11	M7 / Q11	M8 / Q12
Focus	**Technical Feasibility**	**Material Compatibility**	**Performance (GPU Form)**	**Server Proof & System Availability Model**	**Component Reliability for System Model**	**Partial Rack & System Availability**	**Workshop**	**Technology to Market Strategy**
Design & Performance	**Concept Design Development risk retired** (channel, layers, integration)	**2 materials sets**	**SiCP on TTV (>1.5kW) Rth < 0.015 K/W**	**SiCP on Server (80 KW/m3) Rth <=0.015 K/W**		**SiCP on server in Partial Rack Rth <0.01 K/W Cooling Power < 3% IT**		
Modelling	**Model Rth <=0.01 K/W**			Path to Rth <=0.01 K/W		Models at 80kW/m3		
Microfluidics	MCM Process **Build GPU Proxy**	Build coupons	Build	Build	Build	Build		
Microbond Integration	**TIM Material(s) with Rth <=0.005 K/W** and Process	Build coupons **2 materials sets**	Build with proxy TIM	Build	Build	Build		
Coolant Delivery	CDM Process	Build coupons	Build	Build	Build	Build		
Reliability	**Identify top risk categories**	**FMEA/Test Plan Material Comp 2wk** Coolants		**Physics of failure >100hrs Model >95%** Path to Sys 99.982%	**Component Reliability (MTTF)** System Model >99.98%	**>250hrs System 99.982%**		
Tech2Market				**ROI path <7yr IRR path >=10%**			**Workshop >=20 attendees**	**1st Market Application**
System	*GPU Proxy Spec Integration Plan Cooling Power <3% IT*	*System Reliability Test Plan FMEA CDU*	***TTV & Power maps*** *Server non-uniform power model (>3kW/U)*	***Server Test Cooling Power <3% IT***	*System RBD model*	***Partial Rack (>10KW) System Test System <5% IT load*** *Model System Avail*		

arpa·e
CHANGING WHAT'S POSSIBLE

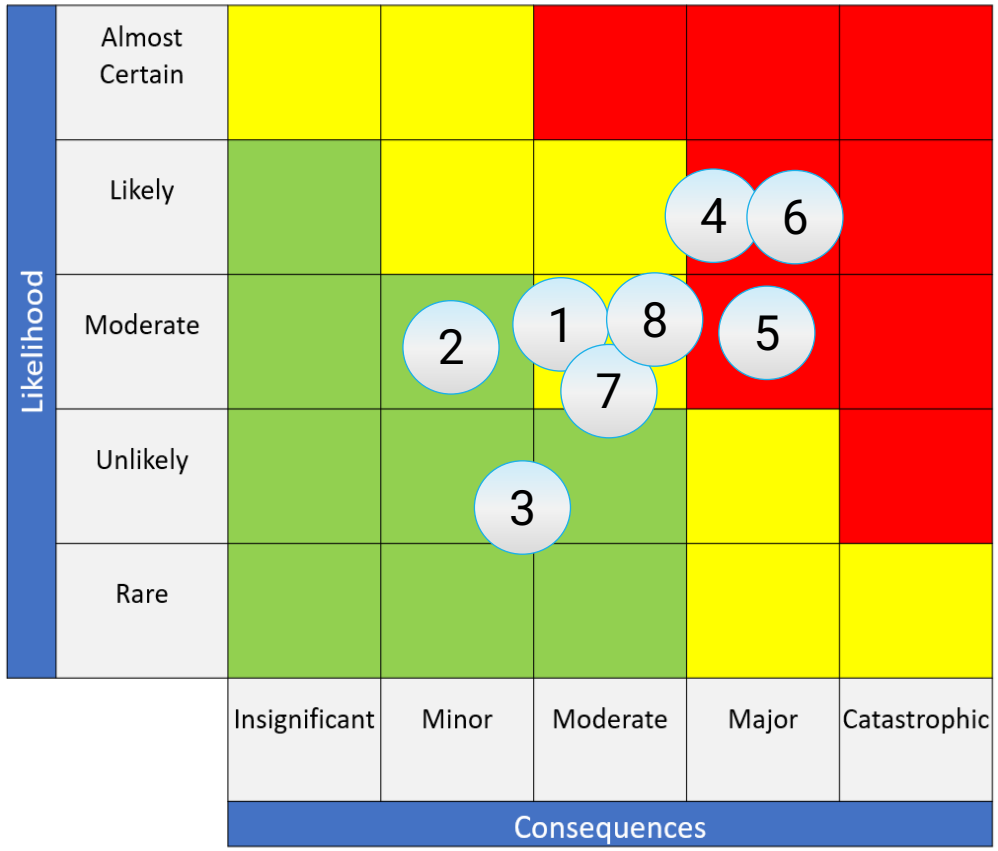
Challenges and Risks

- Key risks are driven primarily by reliability considerations as *microfluidic silicon cold plates (SiCP) have never been attempted at scale for chip cooling*



Challenges and Risks

- Key risks are driven primarily by reliability considerations as *microfluidic silicon cold plates (SiCP) have never been attempted at scale for chip cooling*



#	Risk
1	Insufficient thermal conductivity of SiCP to GPU interface causes > 0.01 K/W
2	Thermal non-uniformity in SiCP causes hot spots and < 1.5kW TDP
3	Pressure drop too high resulting in CDU power requirements exceeding IT budget target of < 3%
4	Structure failure between SiCP and GPU
5	Leak in the manifold or fluid connections
6	Materials compatibility with coolant
7	Cost prohibitively high
8	Slow adoption of technology due to cost or performance mismatch

arpa·e
CHANGING WHAT'S POSSIBLE