Hyperion Research AI Research Update:

What's Going On Around The World, And Our Research Plans For Studying AI For Science

Earl Joseph
ejoseph@hyperionres.com

Visit Our Website: www.HyperionResearch.com

Twitter: HPC_Hyperion@HPC_Hyperion

Hyperion Research helps organizations make effective decisions and seize growth opportunities by providing research and recommendations in both high performance computing and emerging technology areas.

Hyperion Research HPC Activities

- **Track all HPC servers sold each quarter**
  - By 28 countries
- **4 HPC User Forum meetings each year**
- **Publish 85 plus research reports each year**
- **Visit all major supercomputer sites & write reports**
- **Assist in collaborations between buyers/users and vendors**
- **Assist governments in HPC plans, strategies and direction**
- **Maintain 5 year forecasts in many areas/topics**
- **Develop a worldwide ROI measurement system**
- **AI-HPDA program and tracking**
- **HPC Cloud usage tracking**
- **Cyber Security**
- **Quantum Computing**
- **Mapping applications to algorithms to architectures**

Agenda

1. **Some Interesting Findings From Our Studies**
2. **Chinese Plans and Activities**
3. **European Plans and Activities**
4. **Our Plans For Researching AI For Science: Key Questions To Be Studied**
5. **Summary: Some Predictions**

Why AI Is Important To Nations

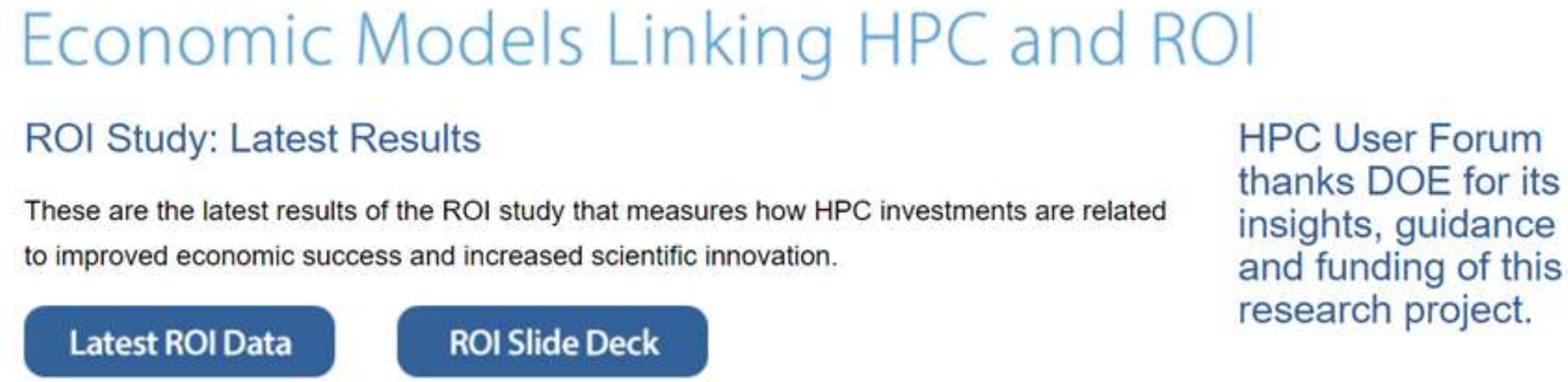
- **It has a major potential for competitive advantage**
  - It has the potential to leap-frog science and other areas
  - Economic value is very high
  - Falling behind could happen very fast, and it will be hard to recover
  - It may determine who owns the "Cloud"
- **It's creating new capabilities, new markets and new ways to quickly solve difficult problems**
  - Precision medicine may be the largest economic area
  - Homeland security, defense, fraud detection are the early areas
  - Automating certain activities will redefine many things, e.g. cyber security, steering experiments, analysis of results, and potentially creating new theories
- **It can help address the scientific labor shortage**
  - Europe and the US have a shortage of scientists and engineers – and need to find ways to make them more productive

Why AI Is Important To Science

- **It adds new research capabilities**
  - Inferencing may become the 4th branch of the scientific method
  - Handle massive, heterogeneous data volumes
  - Help steer modeling and simulation
  - Bypass unproductive areas of problem spaces
  - Enables unique insights
- **It is potentially applicable to every scientific (and engineering) domain**
  - Biology, chemistry-materials science, physics, earth science, space science-astronomy, also humanities/social sciences
  - Not to forget precision medicine, automated driving, cyber security, smart cities, IoT
- **It can help increase scientific productivity**
  - Handle grunt work so researchers can focus on innovation

THE ROI From HPC and AI

www.HyperionResearch.com/roi-with-hpc/

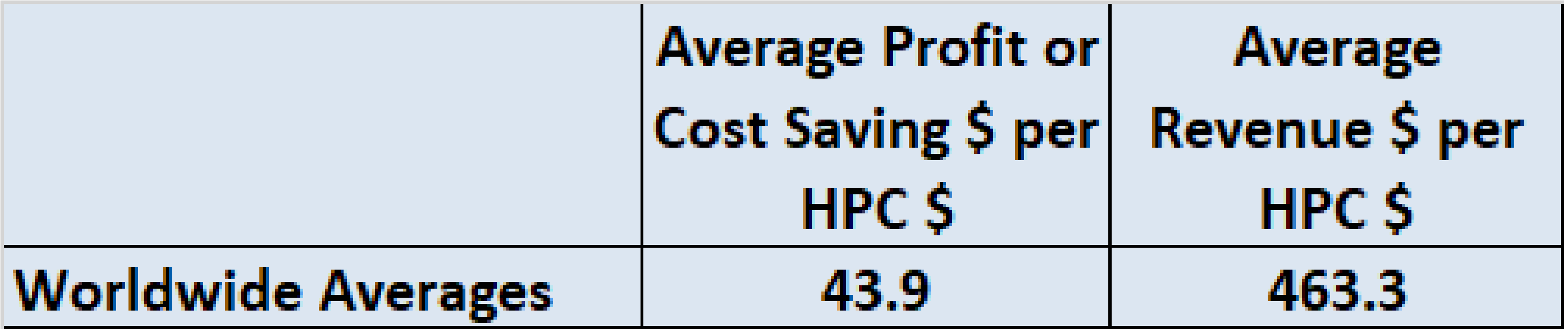


Economic Models Linking HPC and ROI

ROI Study: Latest Results

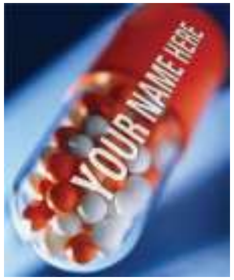
These are the latest results of the ROI study that measures how HPC investments are related to improved economic success and increased scientific innovation.

HPC User Forum thanks DOE for its insights, guidance and funding of this research project.



	Average Profit or Cost Saving $ per HPC $	Average Revenue $ per HPC $
Worldwide Averages	43.9	463.3

The Most Important Use Cases



Precision Medicine

Automated Driving Systems

Fraud and anomaly detection

Affinity Marketing

Business Intelligence

Cyber Security

IoT

High Growth Areas: HPDA-AI

- HPDA is growing faster than overall HPC market
- AI subset is growing faster than all HPDA

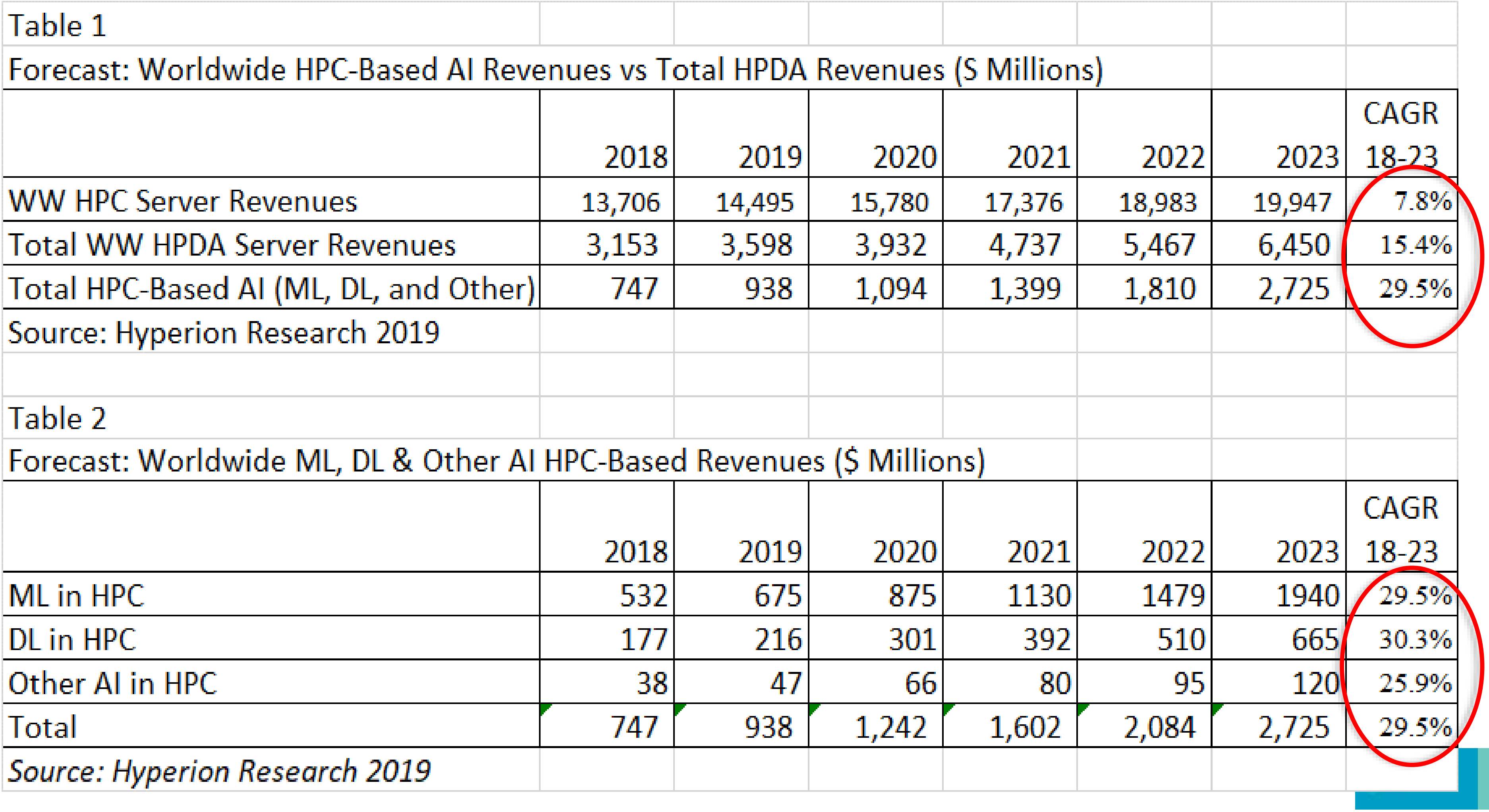


Table 1

Forecast: Worldwide HPC-Based AI Revenues vs Total HPDA Revenues (S Millions)

	2018	2019	2020	2021	2022	2023	CAGR 18-23
WW HPC Server Revenues	13,706	14,495	15,780	17,376	18,983	19,947	7.8%
Total WW HPDA Server Revenues	3,153	3,598	3,932	4,737	5,467	6,450	15.4%
Total HPC-Based AI (ML, DL, and Other)	747	938	1,094	1,399	1,810	2,725	29.5%

Source: Hyperion Research 2019

Table 2

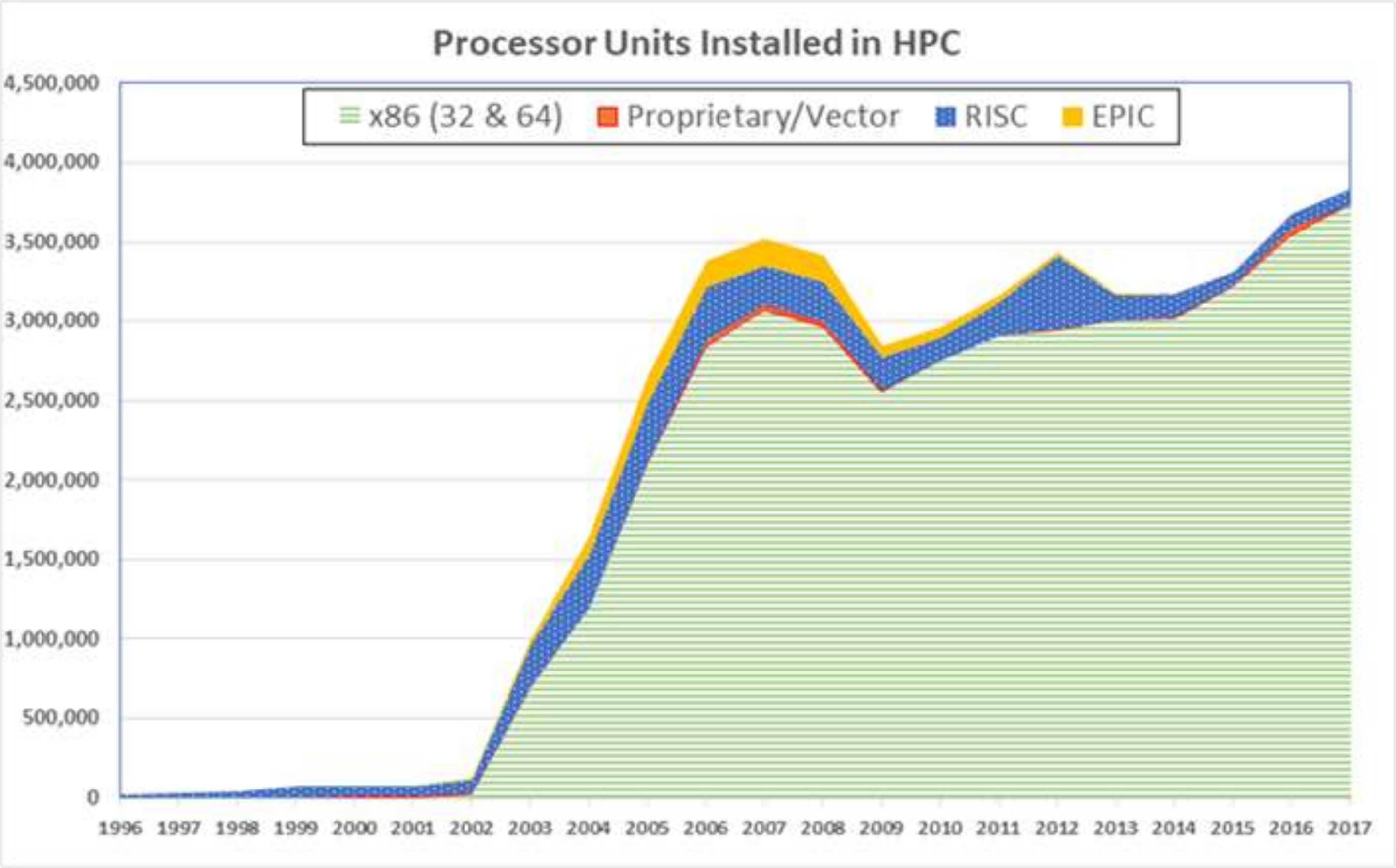
Forecast: Worldwide ML, DL & Other AI HPC-Based Revenues ($ Millions)

	2018	2019	2020	2021	2022	2023	CAGR 18-23
ML in HPC	532	675	875	1130	1479	1940	29.5%
DL in HPC	177	216	301	392	510	665	30.3%
Other AI in HPC	38	47	66	80	95	120	25.9%
Total	747	938	1,242	1,602	2,084	2,725	29.5%

Source: Hyperion Research 2019

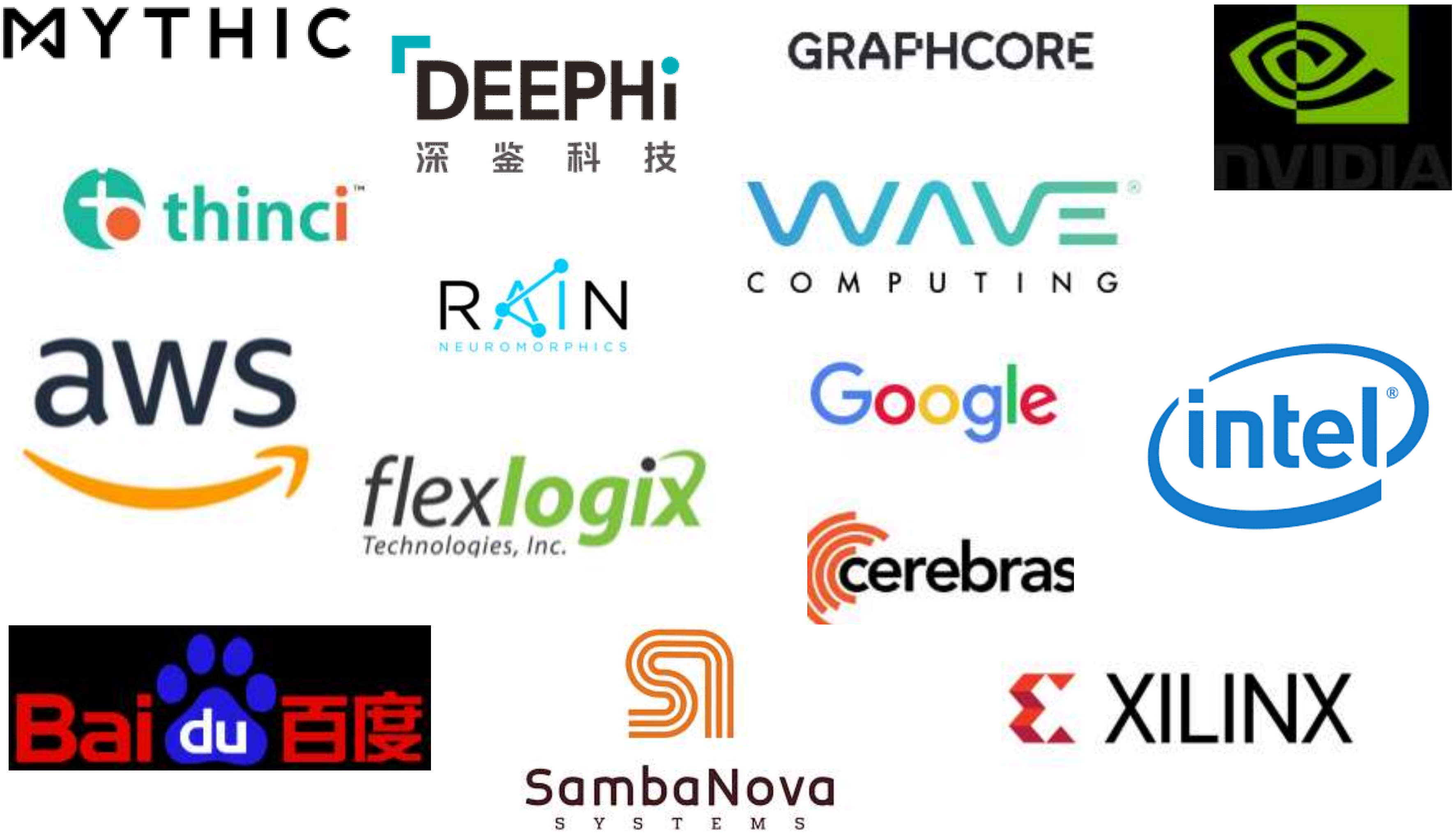
Tipping Points: How Quickly Buyers Can Change (AI Could Happen This Way)

Processor Units Installed in HPC from 1996 to 2017



Source: Hyperion Research, 2018

Emergence of AI-Specific Hardware Ecosystem



AI-HPDA Algorithm Report: Mapping Algorithms to Verticals & System Requirements

https://hyperionresearch.com/proceed-to-download/?doctodown=hpda-algorithm-report

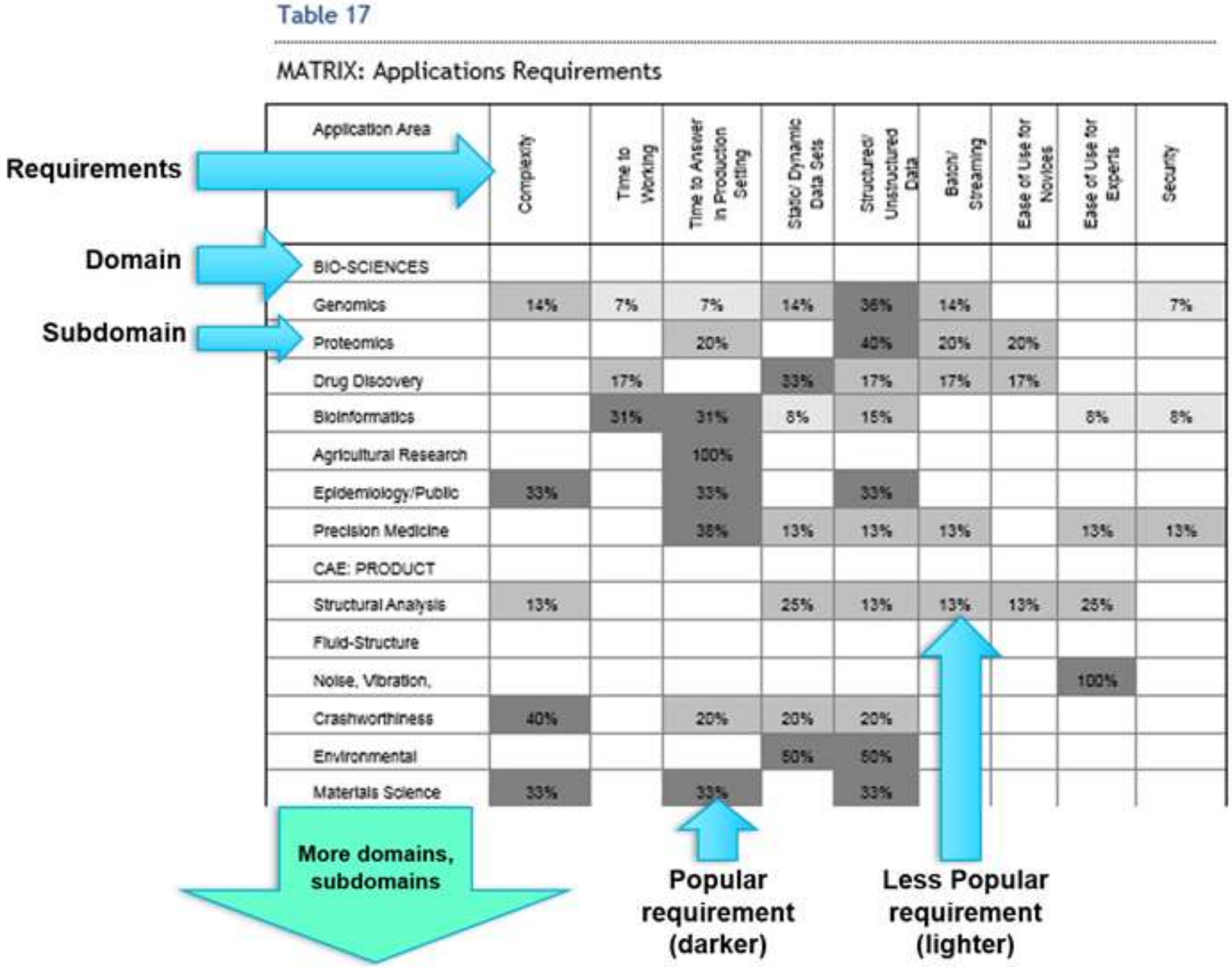


Table 17

MATRIX: Applications Requirements

Application Area	Complexity	Time to Working	Time to Answer in Production Setting	Static/ Dynamic Data Sets	Structured/ Unstructured Data	Batch/ Streaming	Ease of Use for Novices	Ease of Use for Experts	Security
BIO-SCIENCES									
Genomics	14%	7%	7%	14%	36%	14%			7%
Proteomics			20%		40%	20%	20%		
Drug Discovery		17%		33%	17%	17%	17%		
Bioinformatics		31%	31%	8%	15%			8%	8%
Agricultural Research			100%						
Epidemiology/Public	33%		33%		33%				
Precision Medicine			38%	13%	13%	13%		13%	13%
CAE: PRODUCT									
Structural Analysis	13%			25%	13%	13%	13%	25%	
Fluid-Structure									
Noise, Vibration,								100%	
Crashworthiness	40%		20%	20%	20%				
Environmental				50%	50%				
Materials Science	33%		33%		33%				

Requirements

Domain

Subdomain

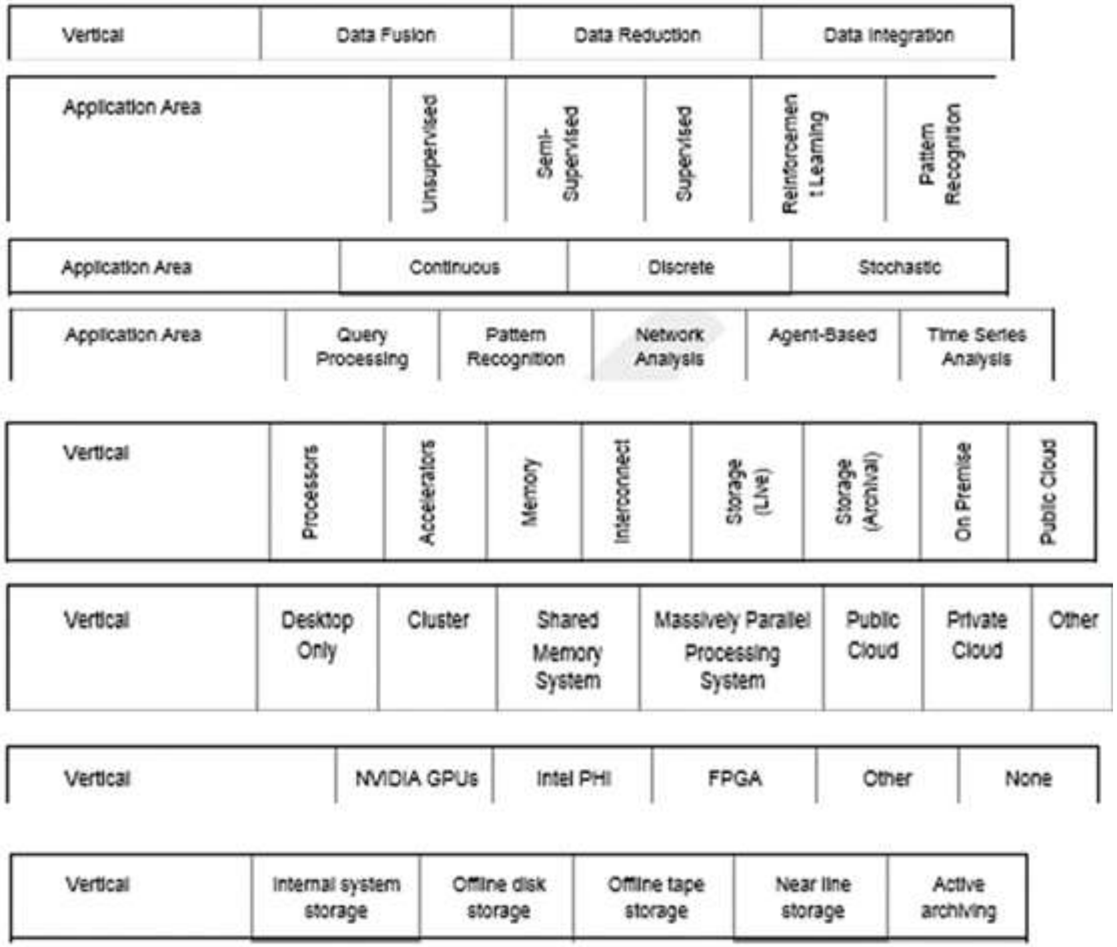
More domains, subdomains

Popular requirement (darker)

Less Popular requirement (lighter)

AI- HPDA Algorithm Report: Mapping Algorithms to Verticals & System Requirements

https://hyperionresearch.com/proceed-to-download/?doctodown=hpda-algorithm-report



- **Data Ingestion**

Vertical	Data Fusion	Data Reduction	Data Integration

- **Machine Learning**

Application Area	Unsupervised	Semi-Supervised	Supervised	Reinforcemen t Learning	Pattern Recognition

- **Numeric Optimization**

Application Area	Continuous	Discrete	Stochastic

- **Data Mining and Simulation**

Application Area	Query Processing	Pattern Recognition	Network Analysis	Agent-Based	Time Series Analysis

- **Hardware Requirements**

Vertical	Processors	Accelerators	Memory	Interconnect	Storage (Live)	Storage (Archival)	On Premise	Public Cloud

- **System Architecture Requirements**

Vertical	Desktop Only	Cluster	Shared Memory System	Massively Parallel Processing System	Public Cloud	Private Cloud	Other

- **Accelerator Requirements**

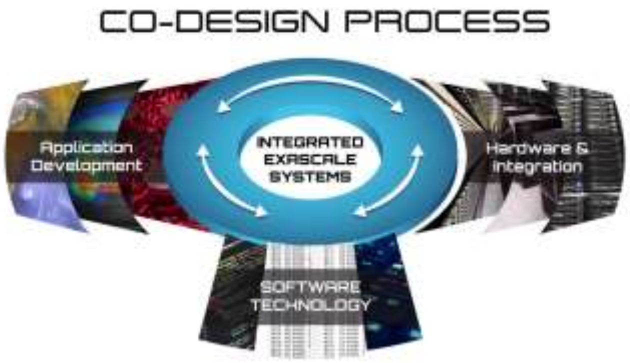
Vertical	NVIDIA GPUs	Intel PHI	FPGA	Other	None

- **Storage Requirements**

Vertical	Internal system storage	Offline disk storage	Offline tape storage	Near line storage	Active archiving

Co-Design

- **AI chips will be centered on co-design, with specific tasks in mind. Examples:**
  - Low-power ASICs at the edge
  - Custom AI chips in hyperscale data centers or the cloud
- **GPUs will remain important but not for all AI workloads.**
- **Software and model-designed hardware is the direction forward.**

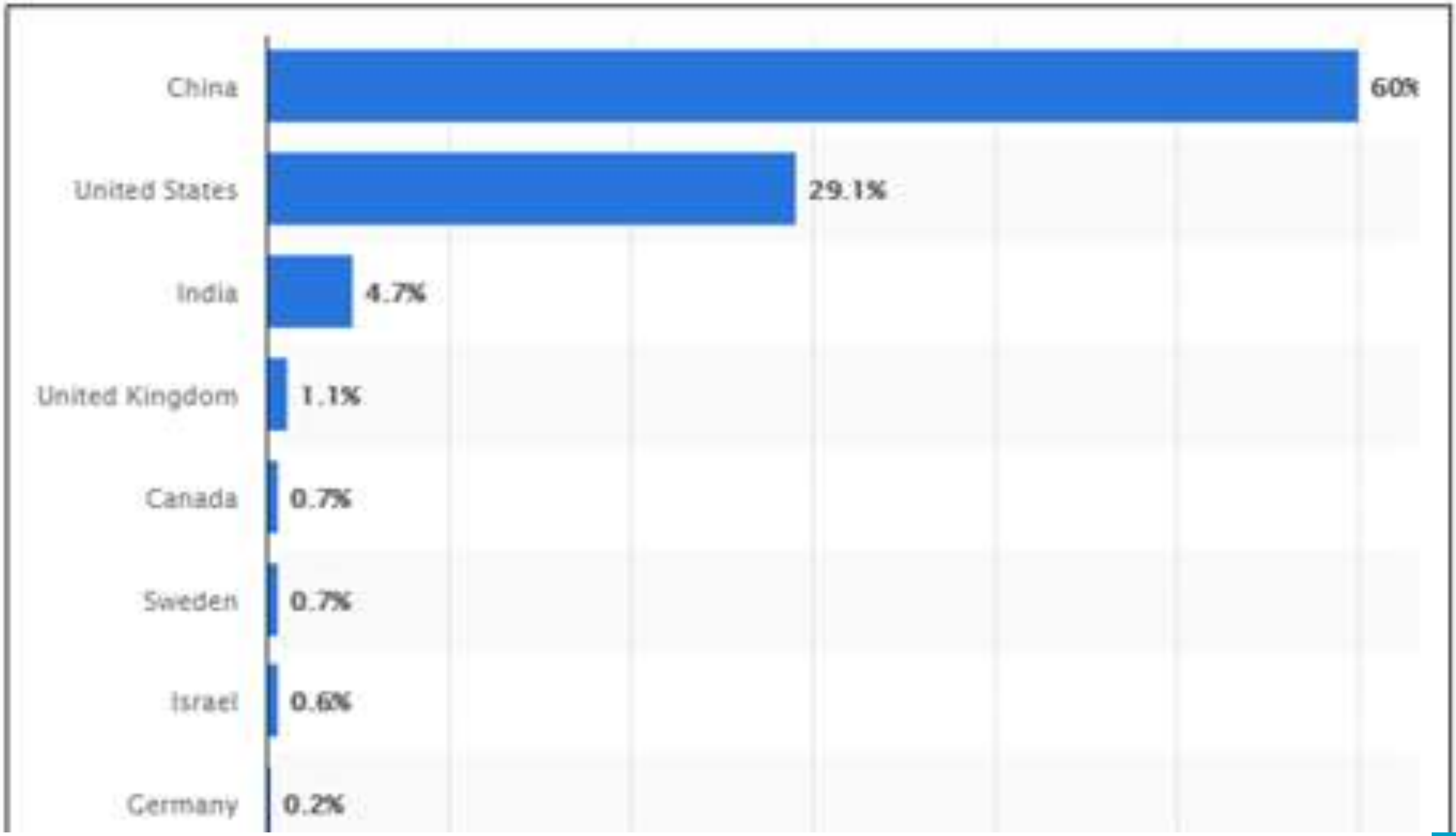


AI Plans And Activities Around The World

AI Investments Around The World

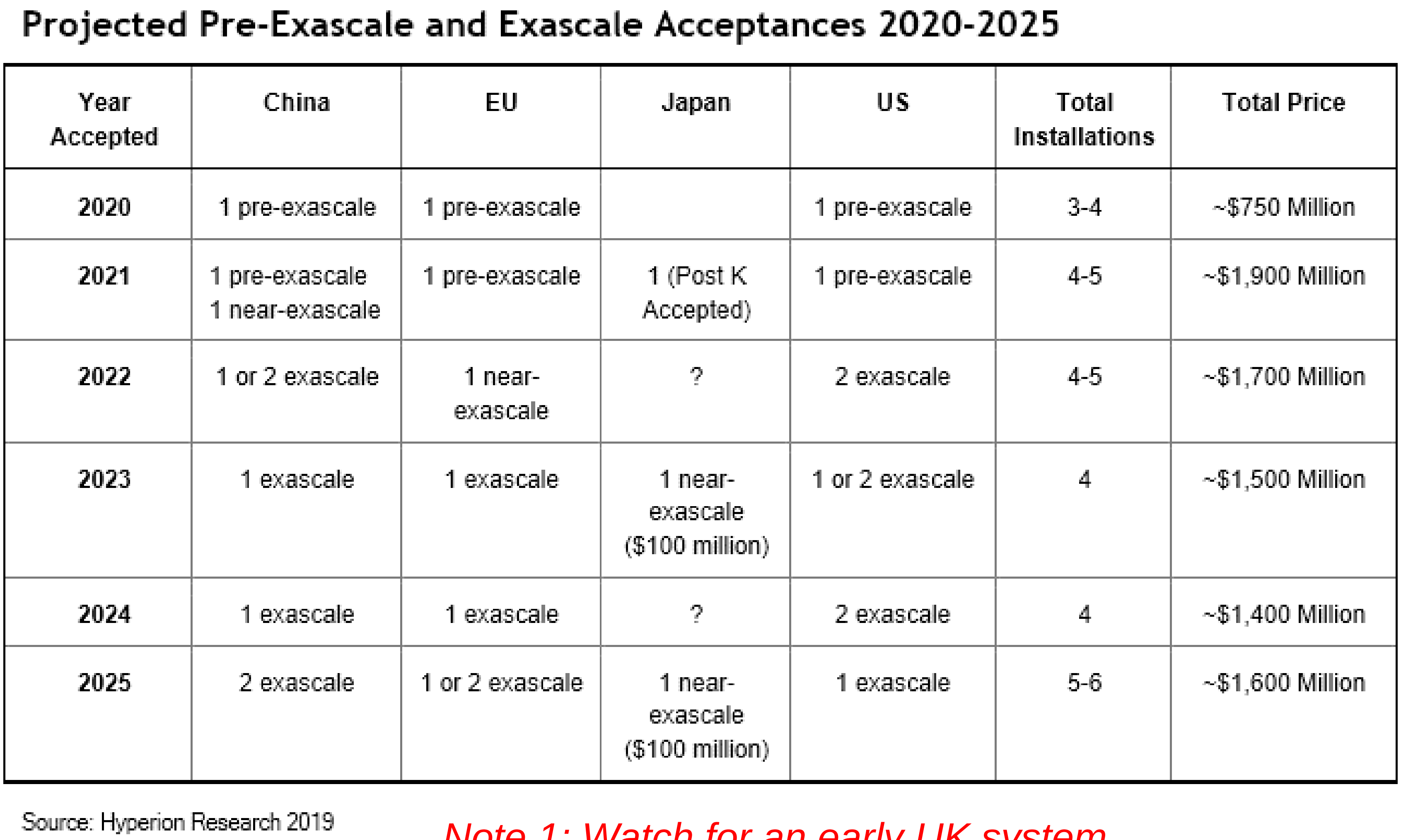
Share of global artificial intelligence (AI) investment and financing by country from 2013 to 1Q'18

https://www.statista.com/statistics/941446/ai-investment-and-funding-share-by-country/



© Hyperion Research

Our Forecast On When & Where Exascale Systems Will Be Installed – Most Now Include AI



Projected Pre-Exascale and Exascale Acceptances 2020-2025

Year Accepted	China	EU	Japan	US	Total Installations	Total Price
2020	1 pre-exascale	1 pre-exascale		1 pre-exascale	3-4	~$750 Million
2021	1 pre-exascale 1 near-exascale	1 pre-exascale	1 (Post K Accepted)	1 pre-exascale	4-5	~$1,900 Million
2022	1 or 2 exascale	1 near-exascale	?	2 exascale	4-5	~$1,700 Million
2023	1 exascale	1 exascale	1 near-exascale ($100 million)	1 or 2 exascale	4	~$1,500 Million
2024	1 exascale	1 exascale	?	2 exascale	4	~$1,400 Million
2025	2 exascale	1 or 2 exascale	1 near-exascale ($100 million)	1 exascale	5-6	~$1,600 Million

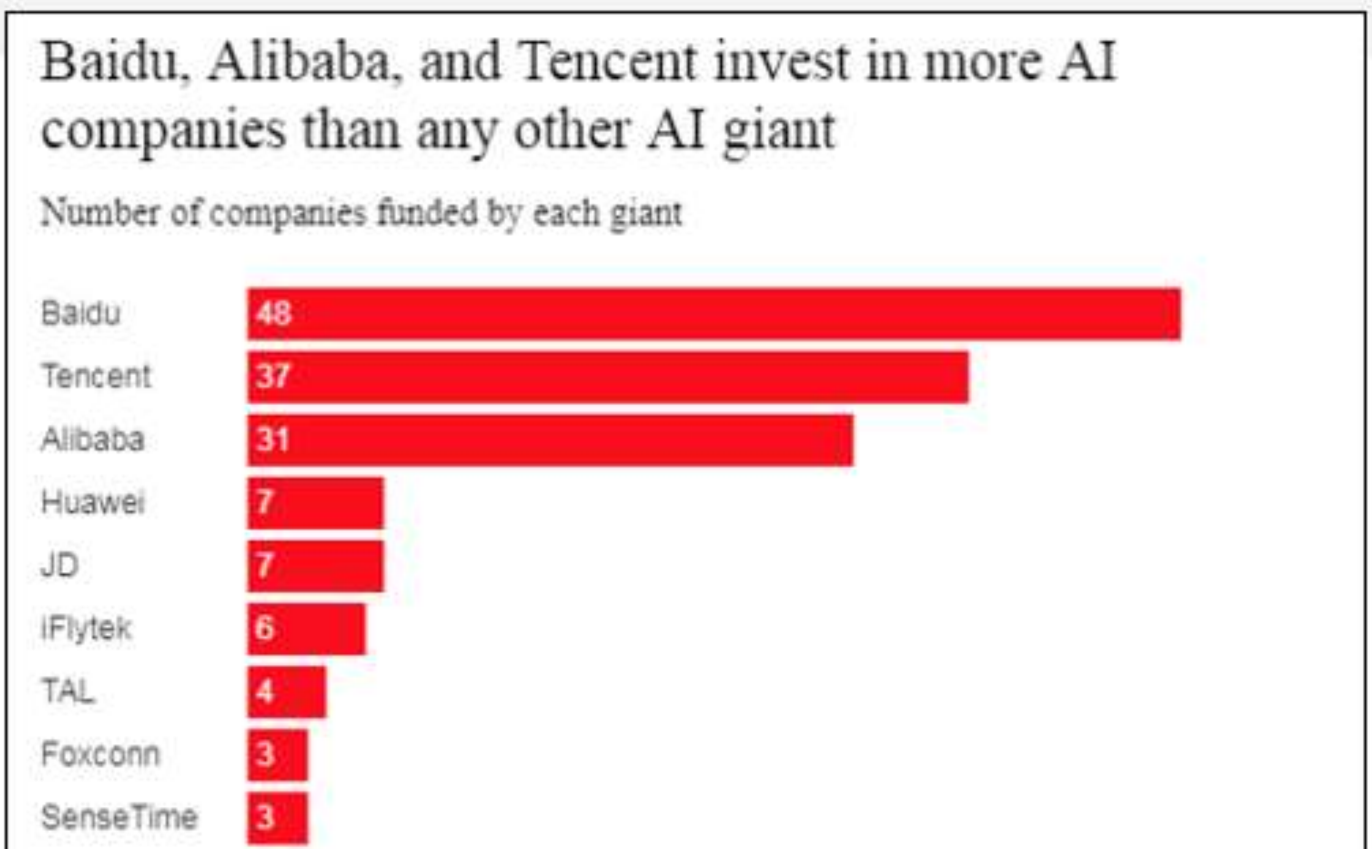
Source: Hyperion Research 2019

Note 1: Watch for an early UK system
Note 2: China may have something in 2020

China Plans And Activities

China AI Activities: CSPs Are Driving Investments In AI

- **More than half of the country’s major AI players have funding ties that lead back to Baidu, Alibaba, and Tencent**
  - From: https://www.technologyreview.com/s/612813/the-future-of-chinas-ai-industry-is-in-the-hands-of-just-three-companies/



Baidu, Alibaba, and Tencent invest in more AI companies than any other AI giant

Number of companies funded by each giant

A Different Take on the AI Startup Ecosystem

- **In the US:**
  - The sentiment with many of the AI HW startups is that each company can find their niche, within their specialty area, and win at just that application, whether it is image processing or NLP or some other AI application.
- **In China:**
  - The trend among the companies is that there will be a few “winners” or successful companies, and the rest will fade away out of the market.

Baidu's View Of The World

- **Their Prediction: by 2020, 70% of servers will have AI processors.**
- **Baidu Kunlun, XPU: AI processors that is general and flexible, power efficient, and has high computing capability.**
- **Built by Samsung, 14nm, 512 Gb/s off-chip memory, 260 TOPS.**
- **Two chips: Kunlun 818-300 (Training) and Kunlun 818-100 (inference).**
- **Many application areas, including speech, NLP, image recognition, ADS, and more.**
- **Chips have been tested in real environments.**

Alibaba, Lingjie Yu, Director of Applied AI

- **Right now there is a trend for heterogeneous computing, and GPUs are not ideal for many workloads as it does not offer true elasticity or multi tenancy.**
- **Inference requires new chips and will be case driven.**
- **China has more AI applications than most other nations.**
- **Software is underinvested right now, and co-design needs to be important to development in AI.**
- **Their advice for startups:**
  - Pick a particular segment for focus, like inference vs. training
  - Don’t compete with the big guys, like NVIDIA
  - Know your niche
  - And the cloud will be the best friend of AI hardware

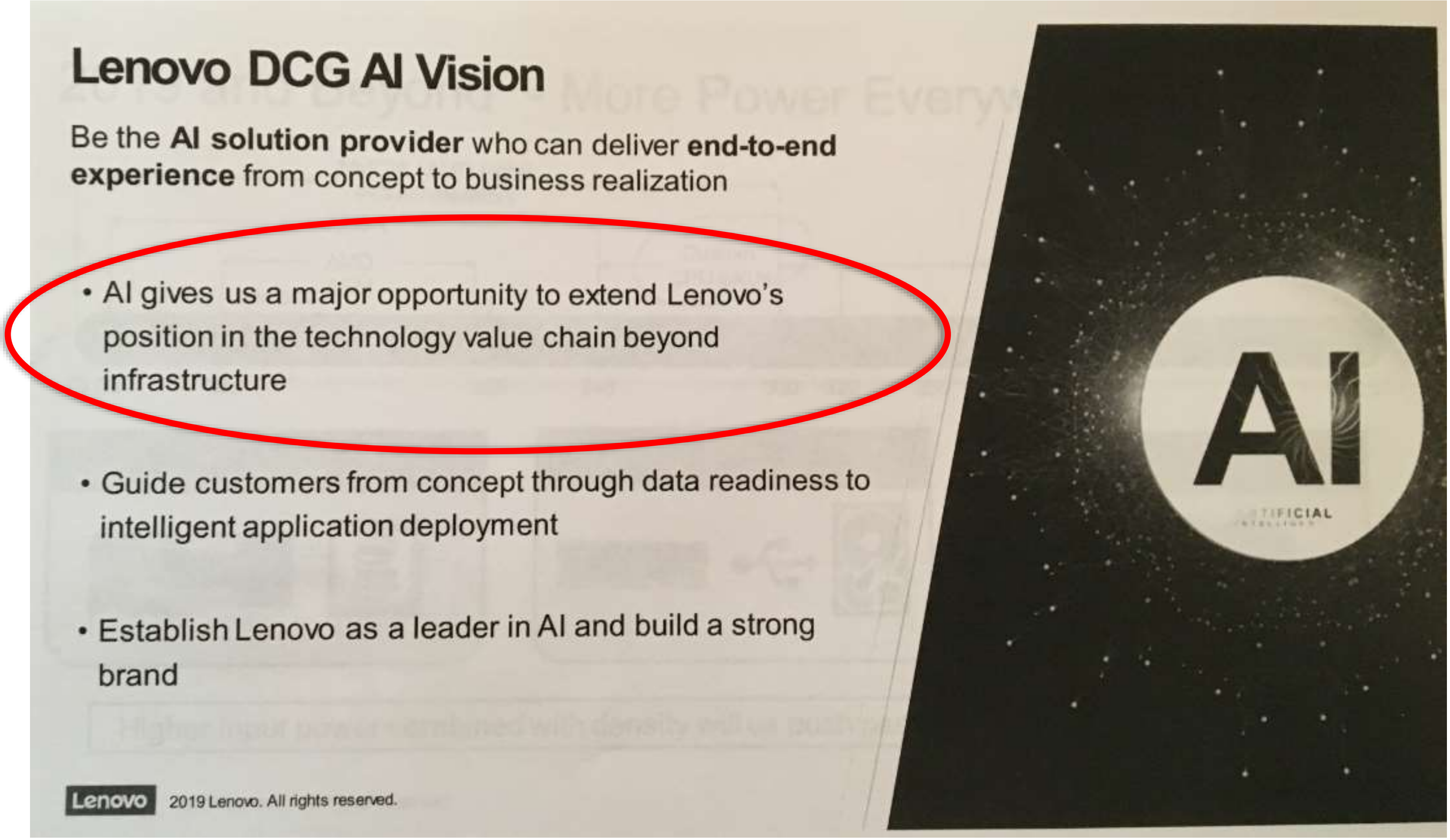
Horizon Robotics, Kai Yu, Founder

- **Horizon just celebrated its 4th year, and was the first mover towards AI smart chips.**
  - Horizon competes in training based apps, not just technology, and competes in total ecosystem.
  - "We are not doing robotics, but rather are developing a horizontal platform for robotics to enable development of autonomous systems (ADS is most exciting right now)."
- **"We care about edge and inference, and we do SW and AI algorithms as well as hardware."**
- **Horizon has 2 product lines, one for ADS (20+ TOPS) and one for smart city video analytics (5TOPS).**
- **Unlike Tesla, which is a black box model, Horizon is an open platform designed to achieve high efficiency by finding the balance between a closed system and an open system.**

Canaan, Zhang Li

- **The K210 AI chip is their main product, an ARM based, RISC-V edge computing AI chip.**
  - They claim it's the first 7nm ASIC.
  - 8mb of RAM on chip.
- **Started development with Bitcoin in mind, and now the 2nd largest blockchain chip manufacturer.**
- **Fabricated at TSMC.**
- **Does audio, visual and 3d rendering, and now has many audio/visual applications like face detection, recognition.**
  - Presently in four main verticals: smart home, industrial sectors, education and agriculture (work with Baidu).
- **5G is crucial for IoT and the middleware is needed to connect the edge to the cloud.**
- **Next generation chip is K510, a 3x improvement over the current K210, which will tape-out at the end of 2019.**

Lenovo Example: Their AI Vision



Lenovo DCG AI Vision

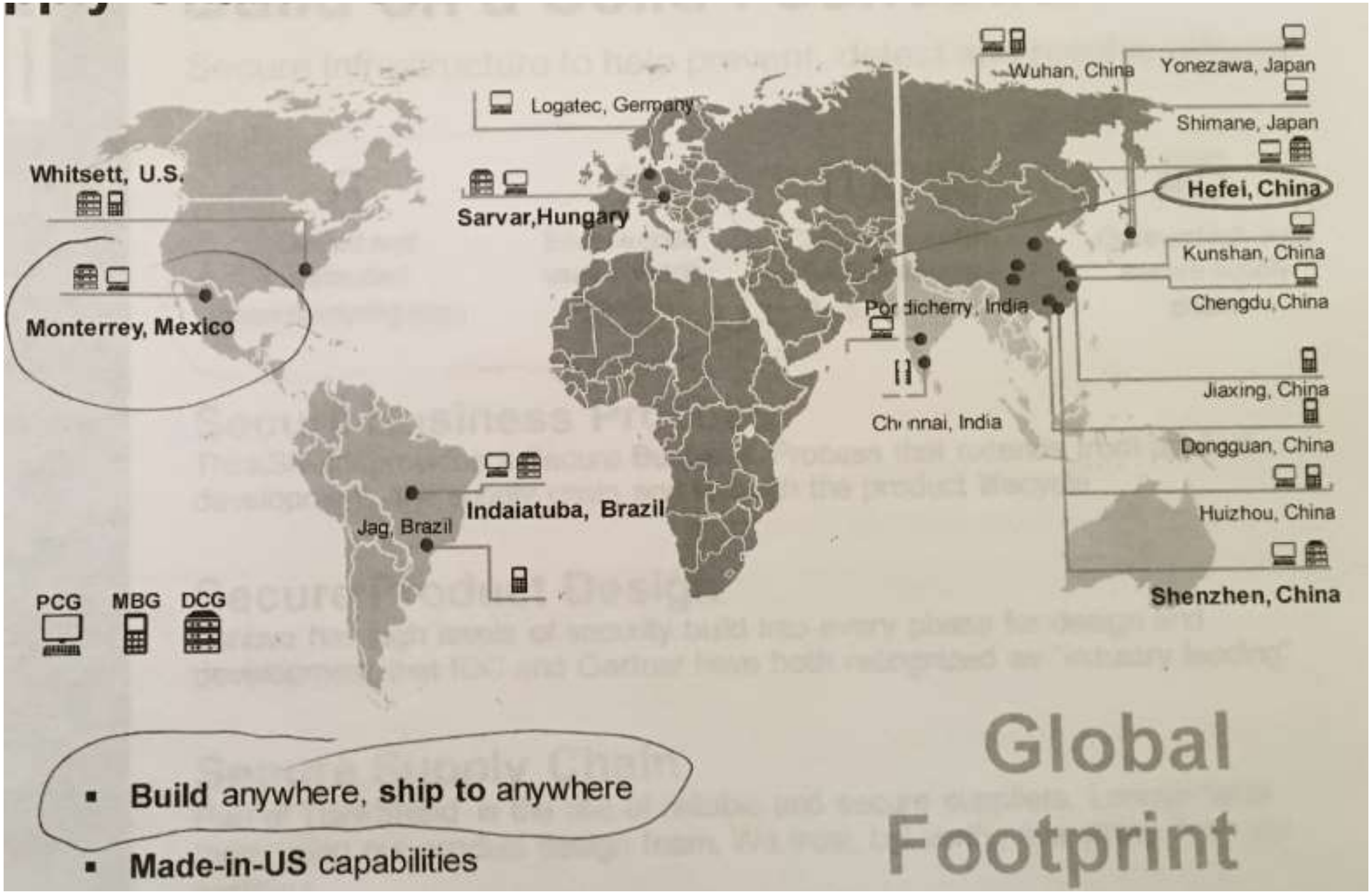
Be the **AI solution provider** who can deliver **end-to-end experience** from concept to business realization

- AI gives us a major opportunity to extend Lenovo’s position in the technology value chain beyond infrastructure
- Guide customers from concept through data readiness to intelligent application deployment
- Establish Lenovo as a leader in AI and build a strong brand

Lenovo 2019 Lenovo. All rights reserved.

Lenovo Example: Ability To Build Regionally and Avoid Tariffs

- **Each factory can switch to building other products**

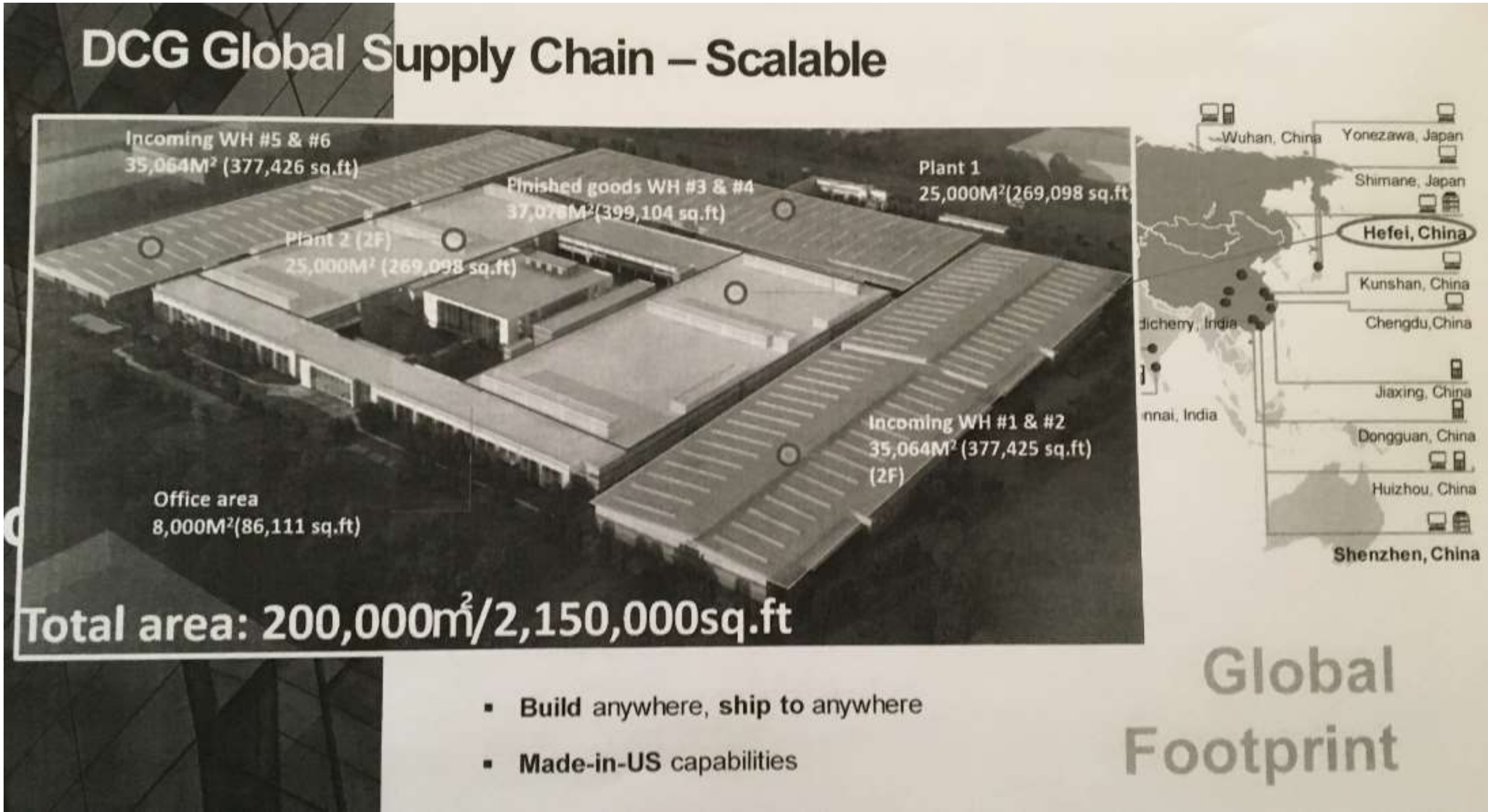


- **Build** anywhere, **ship to** anywhere
- **Made-in-US** capabilities

Global Footprint

Lenovo Example: Full Factory Redesign and Modernization

- **They say it’s the world's largest IT factory**
- **Buying $50 Billion a year in parts**
- ***Will they become the largest computer company?***
- ***Who will they buy next?***



DCG Global Supply Chain – Scalable

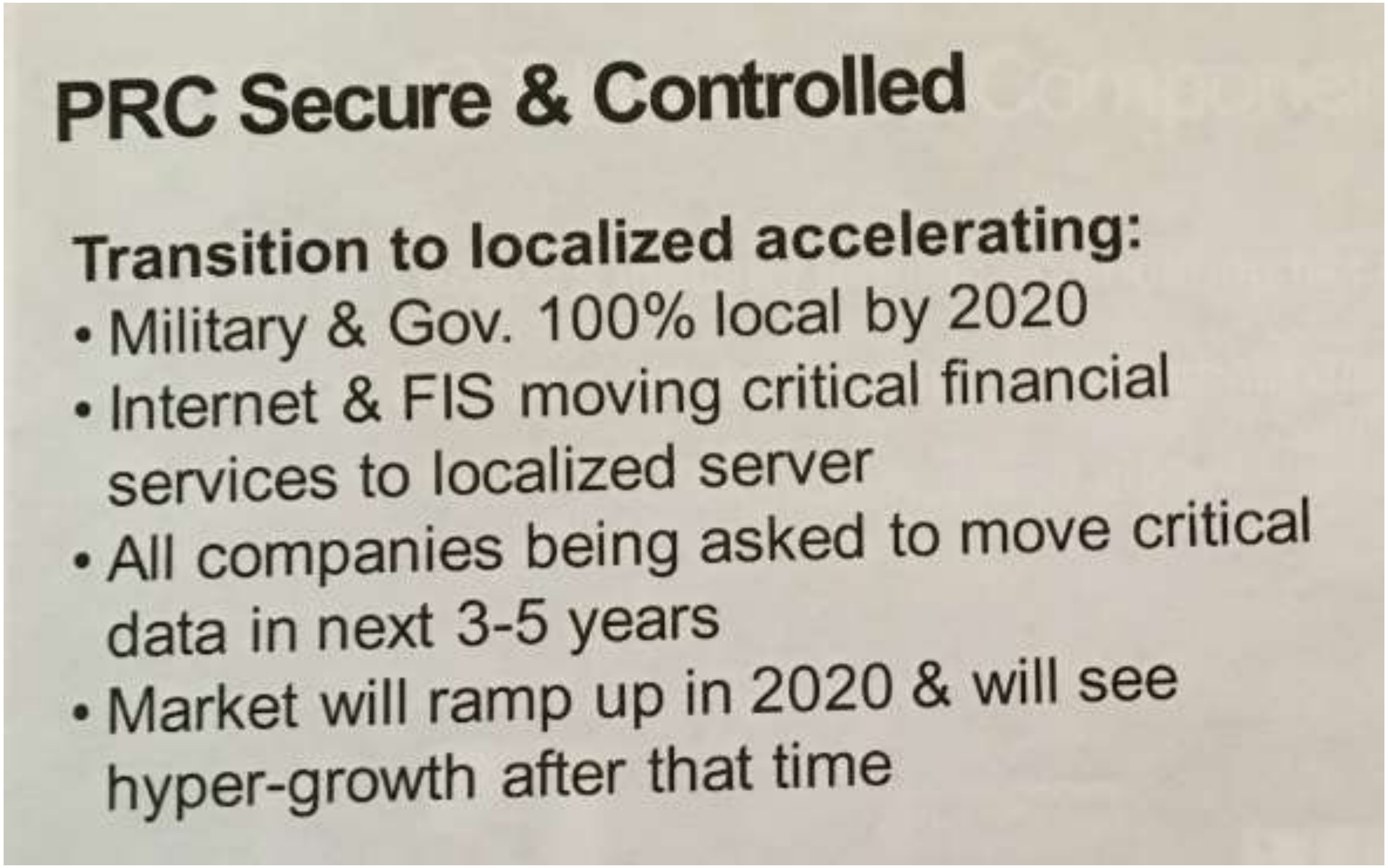
Total area: 200,000m²/2,150,000sq.ft

- Build anywhere, ship to anywhere
- Made-in-US capabilities

Global Footprint

Lenovo Example: China Is Moving To China Built Processors

- **For security and control**



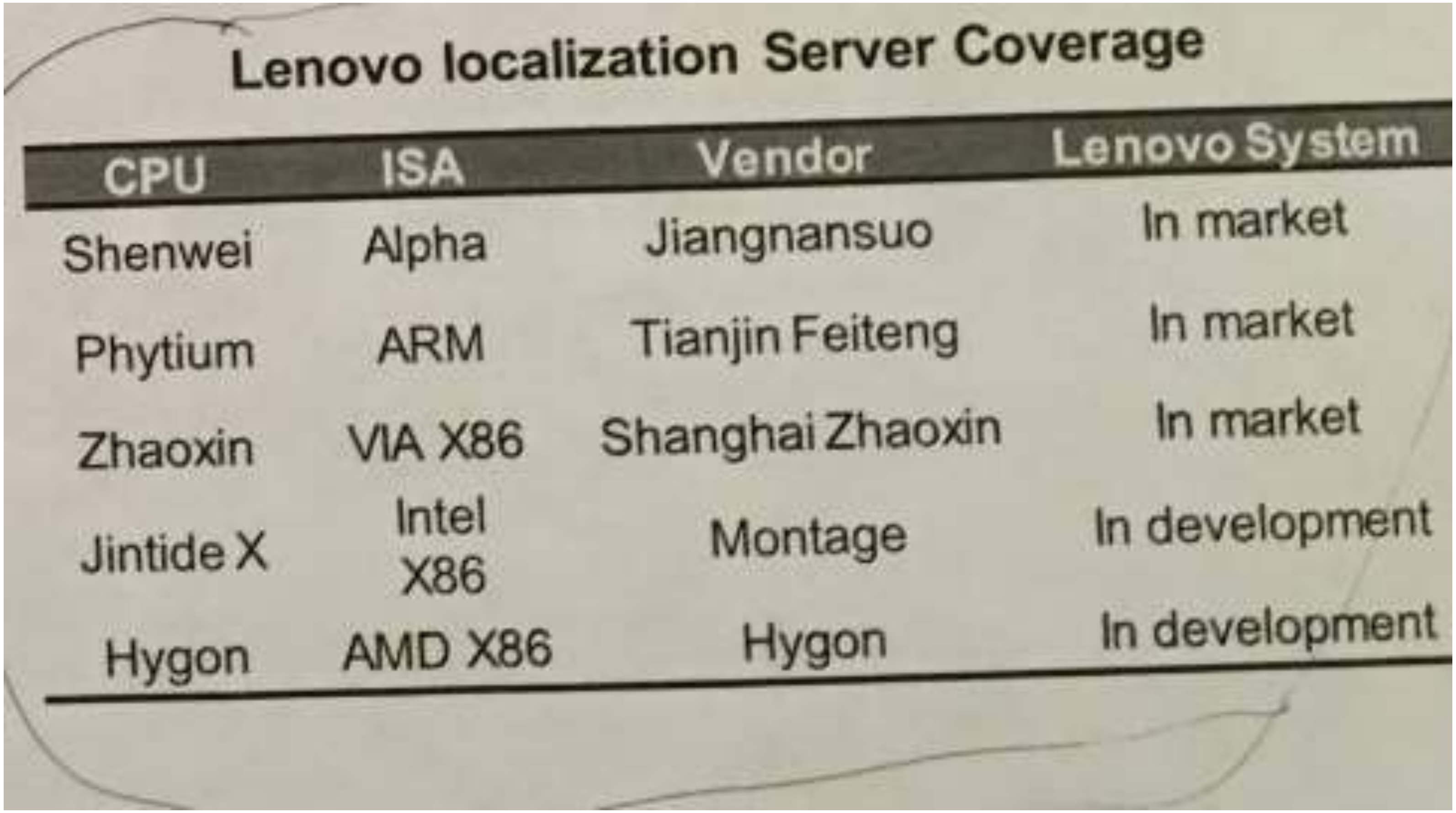
PRC Secure & Controlled

Transition to localized accelerating:

- Military & Gov. 100% local by 2020
- Internet & FIS moving critical financial services to localized server
- All companies being asked to move critical data in next 3-5 years
- Market will ramp up in 2020 & will see hyper-growth after that time

Lenovo Example: … The "Chinese" Processors They Plan To Use

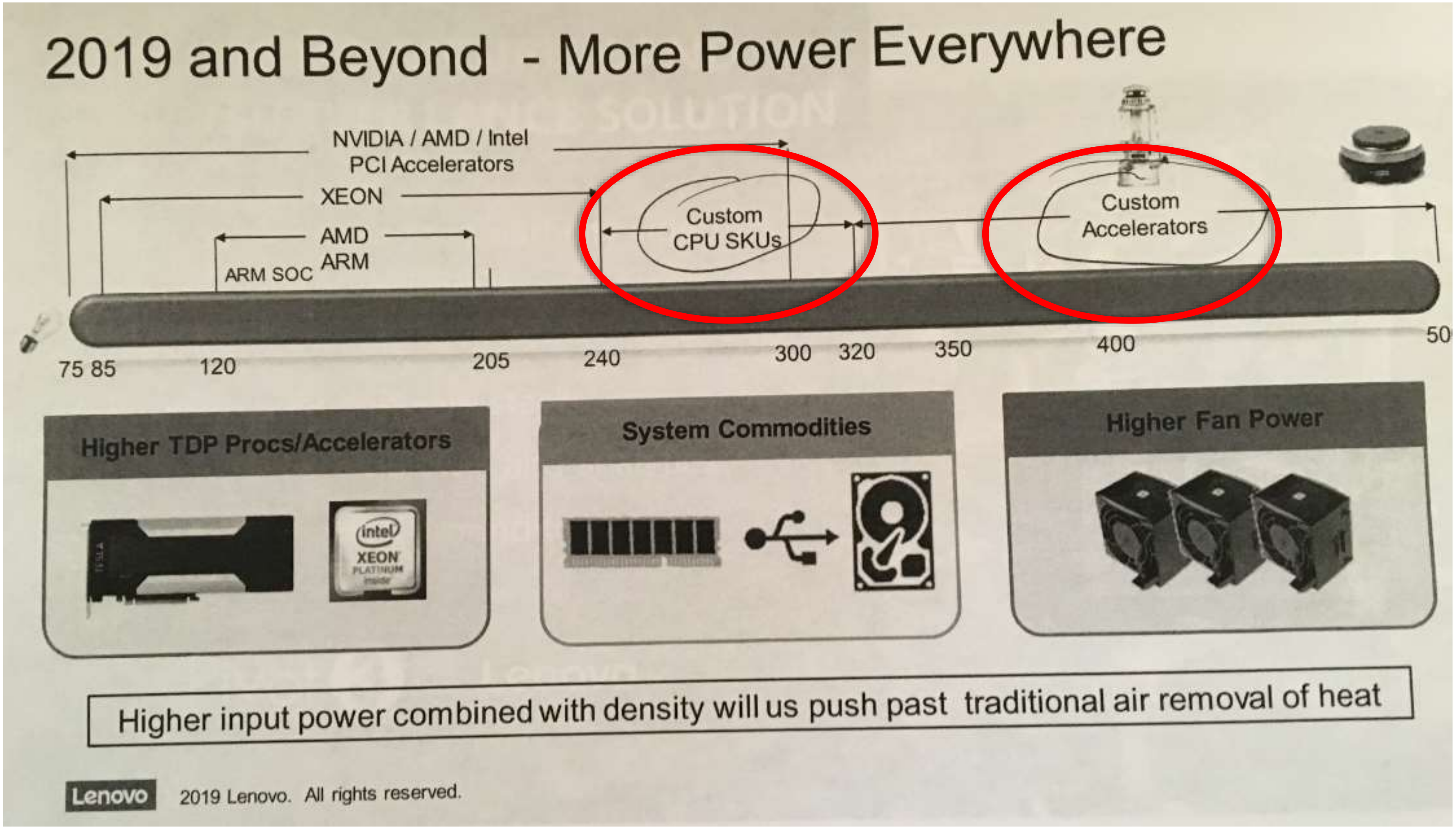
- **Based on x86, Alpha and ARM**
- **Intel and AMD are creating Chinese specific SKUs**



Lenovo localization Server Coverage

CPU	ISA	Vendor	Lenovo System
Shenwei	Alpha	Jiangnansuo	In market
Phytium	ARM	Tianjin Feiteng	In market
Zhaoxin	VIA X86	Shanghai Zhaoxin	In market
Jintide X	Intel X86	Montage	In development
Hygon	AMD X86	Hygon	In development

Lenovo Example: Plans For Custom CPU SKUs, Custom Accelerators & Liquid Cooling



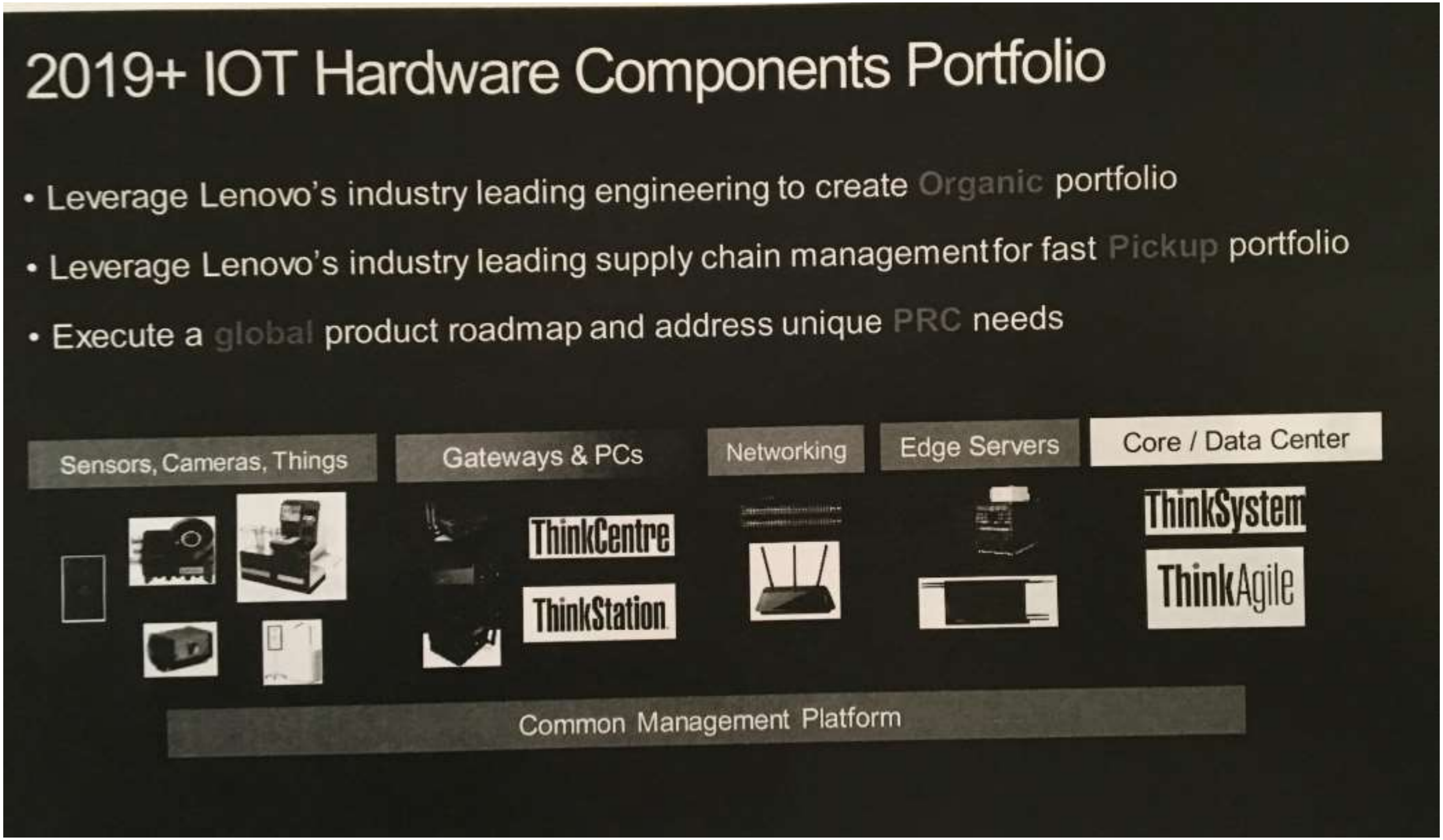
2019 and Beyond - More Power Everywhere

NVIDIA / AMD / Intel
PCI Accelerators
XEON
AMD
ARM
ARM SOC
Custom
CPU SKUs
Custom
Accelerators
75 85
120
205
240
300
320
350
400
50
Higher TDP Procs/Accelerators
intel
XEON
PLATINUM
inside
System Commodities
Higher Fan Power

Higher input power combined with density will us push past traditional air removal of heat

Lenovo 2019 Lenovo. All rights reserved.

Lenovo Example: Building All Types Of IT, Sensors, Cameras, Mini-PCs, etc.



2019+ IOT Hardware Components Portfolio

- Leverage Lenovo's industry leading engineering to create Organic portfolio
- Leverage Lenovo's industry leading supply chain management for fast Pickup portfolio
- Execute a global product roadmap and address unique PRC needs

Sensors, Cameras, Things

Gateways & PCs

ThinkCentre

ThinkStation

Networking

Edge Servers

Core / Data Center

ThinkSystem

ThinkAgile

Common Management Platform

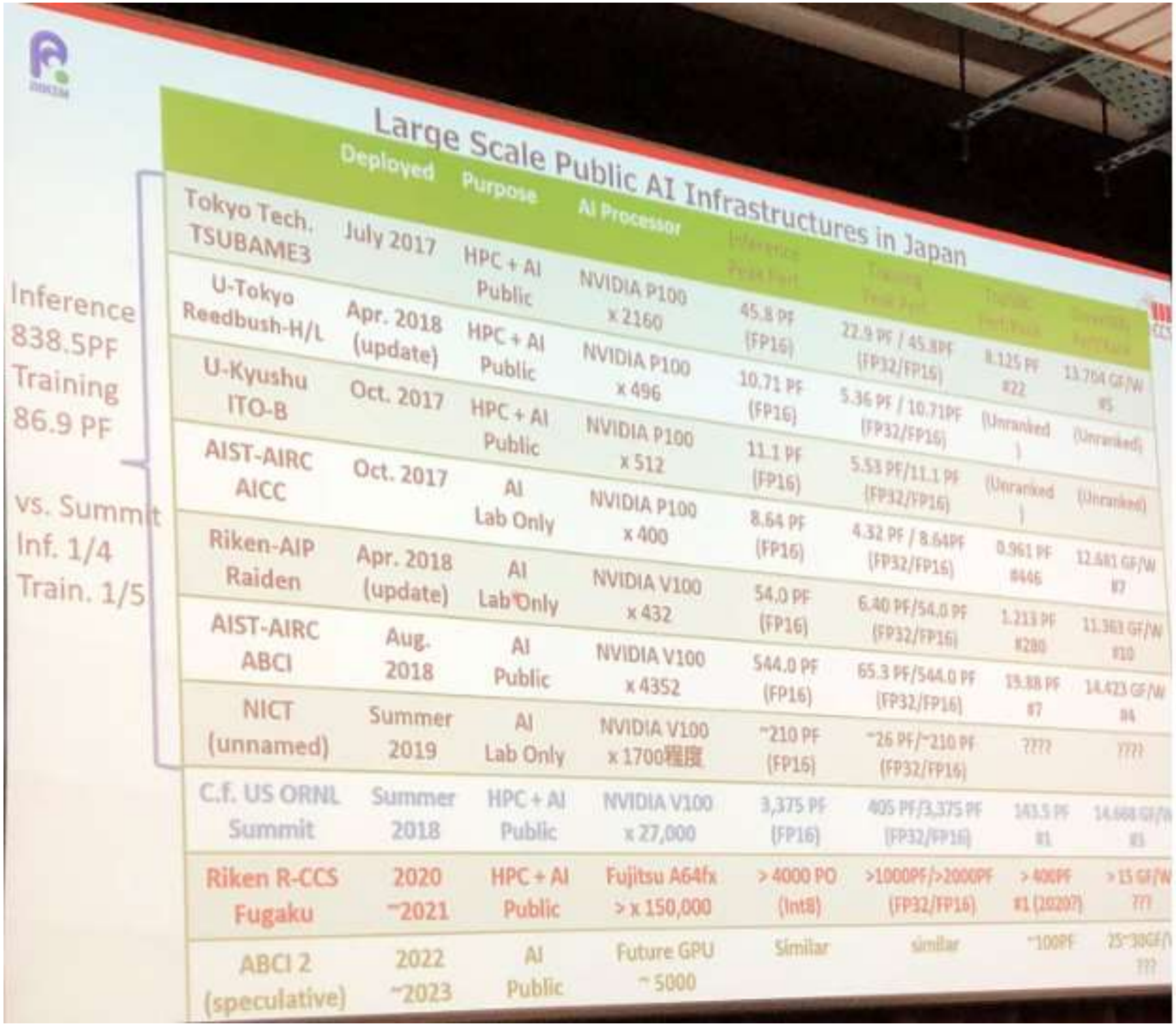
Lenovo Example: Next Steps?

- **Image what would happened if 2 or 3 of these companies merge:**
  - Lenovo
  - Huawei
  - Alibaba
  - Tencent
  - Baidu
- **What if Lenovo buys one or two of these companies?:**
  - Atos/Bull
  - Dell
  - Fujitsu
  - SAP
  - TSMC
  - Erickson
  - Accenture

Chinese Exascale Plans

	Sunway 2020	Sugon Exascale	NUDT 2020
Key User/Developer	Sunway/NRCPC	Sugon/AMD	NUDT
Planned Delivery Date/ Estimated	2020, 4Q (could slip)	2020, 4Q (could slip)	2020, 4Q (could slip)
Planned/Realized Performance (Pflops)	1000	1024	1000
Linpack Performance (PFlops)	600-700	627-732	700-800
Linpack/Peak Performance Ratio (%)	60-70	60-70 (est.)	70-80
High Performance Conjugate Gradient (Pflops/s)	6-7	9.4-10.1	14-16
GF/Watt	30	34.13	20-30
Linpack GF/Watt	20-23	20.9	23.3-32.0

Japanese AI Activities



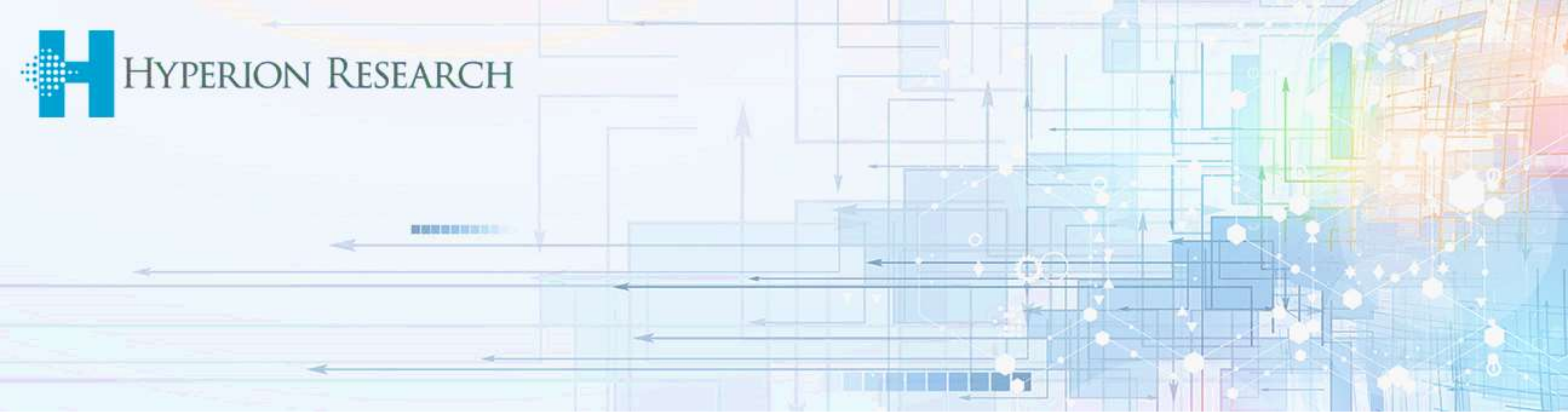
Inference
838.5PF
Training
86.9 PF

vs. Summit
Inf. 1/4
Train. 1/5

Large Scale Public AI Infrastructures in Japan

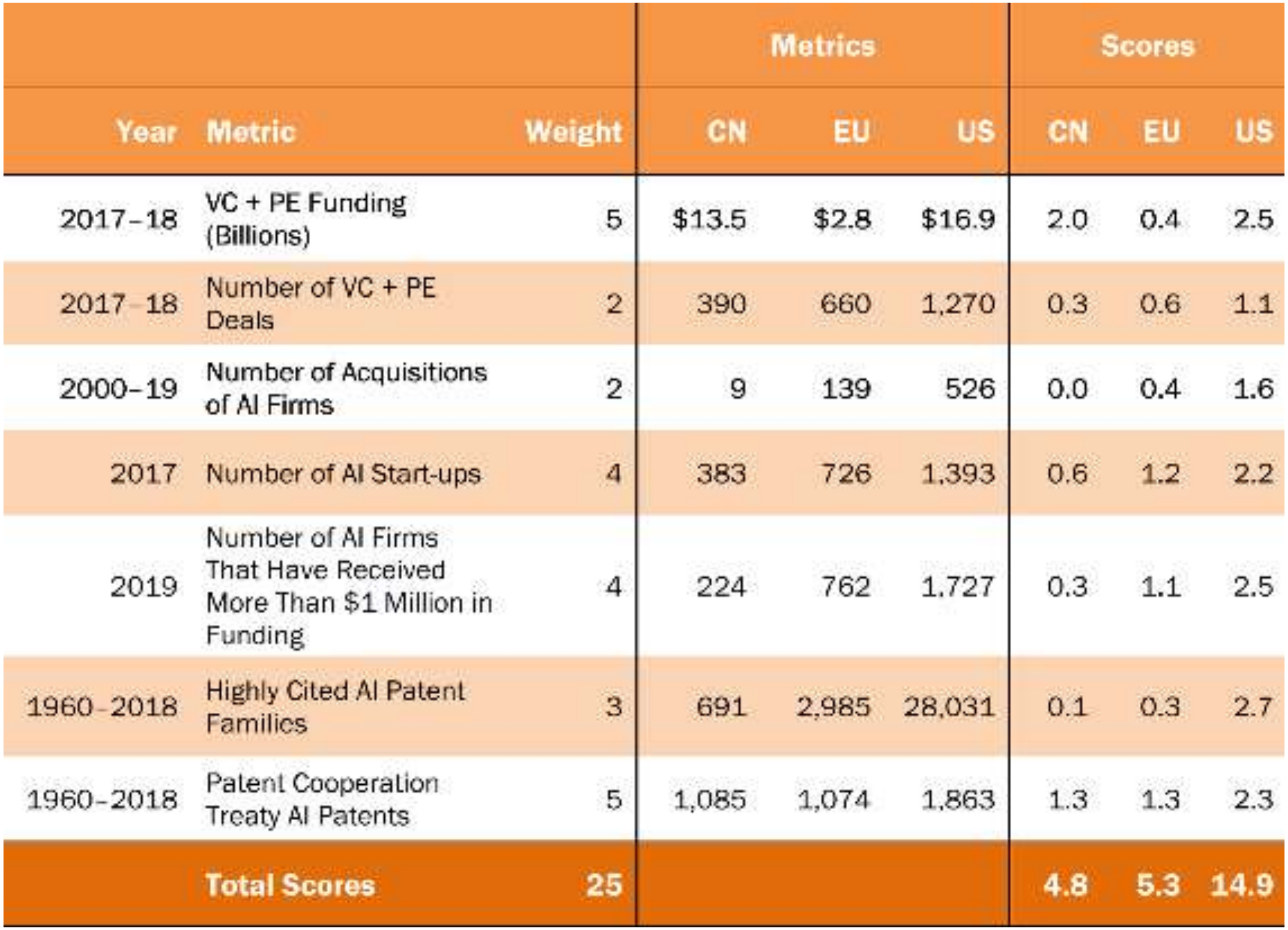
	Deployed	Purpose	AI Processor	Inference Peak Perf.	Training Peak Perf.	Top500 Perf/Rank	Green500 Perf/Rank
Tokyo Tech. TSUBAME3	July 2017	HPC + AI Public	NVIDIA P100 x 2160	45.8 PF (FP16)	22.9 PF / 45.8PF (FP32/FP16)	8.125 PF #22	13.704 GF/W #5
U-Tokyo Reedbush-H/L	Apr. 2018 (update)	HPC + AI Public	NVIDIA P100 x 496	10.71 PF (FP16)	5.36 PF / 10.71PF (FP32/FP16)	(Unranked)	(Unranked)
U-Kyushu ITO-B	Oct. 2017	HPC + AI Public	NVIDIA P100 x 512	11.1 PF (FP16)	5.53 PF/11.1 PF (FP32/FP16)	(Unranked)	(Unranked)
AIST-AIRC AICC	Oct. 2017	AI Lab Only	NVIDIA P100 x 400	8.64 PF (FP16)	4.32 PF / 8.64PF (FP32/FP16)	0.961 PF #446	12.681 GF/W #7
Riken-AIP Raiden	Apr. 2018 (update)	AI Lab Only	NVIDIA V100 x 432	54.0 PF (FP16)	6.40 PF/54.0 PF (FP32/FP16)	1.213 PF #280	11.363 GF/W #10
AIST-AIRC ABCI	Aug. 2018	AI Public	NVIDIA V100 x 4352	544.0 PF (FP16)	65.3 PF/544.0 PF (FP32/FP16)	19.88 PF #7	14.423 GF/W #4
NICT (unnamed)	Summer 2019	AI Lab Only	NVIDIA V100 x 1700程度	~210 PF (FP16)	~26 PF/~210 PF (FP32/FP16)	????	????
C.f. US ORNL Summit	Summer 2018	HPC + AI Public	NVIDIA V100 x 27,000	3,375 PF (FP16)	405 PF/3,375 PF (FP32/FP16)	143.5 PF #1	14.668 GF/W #3
Riken R-CCS Fugaku	2020 ~2021	HPC + AI Public	Fujitsu A64fx > x 150,000	> 4000 PO (Int8)	>1000PF/>2000PF (FP32/FP16)	> 400PF #1 (2020?)	> 15 GF/W ???
ABCI 2 (speculative)	2022 ~2023	AI Public	Future GPU ~ 5000	Similar	similar	~100PF	25~30GF/W ???

© Hyperion Research



EU Plans And Activities

Europe Lags the US, China in AI Private Sector Investment and Patents



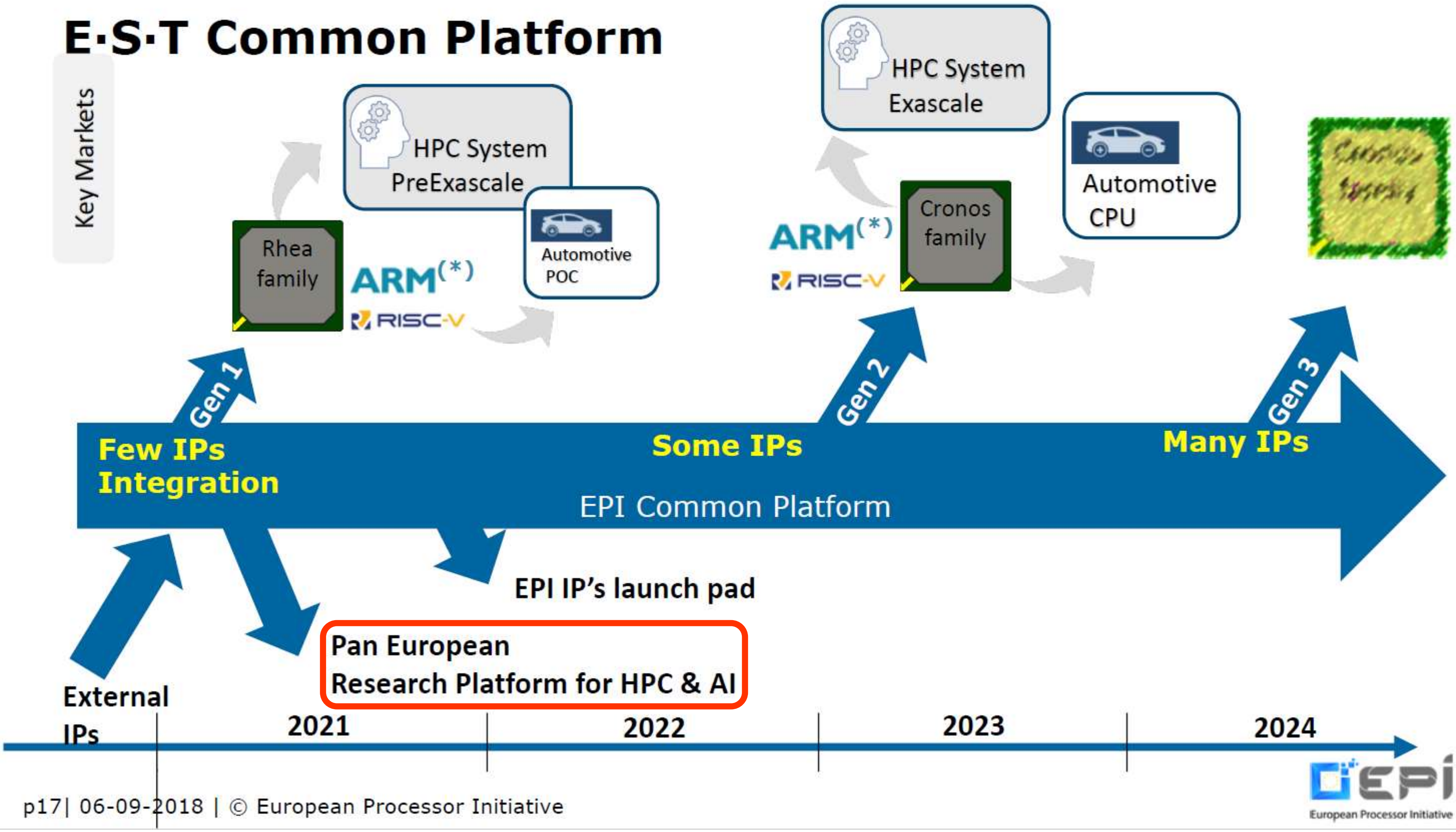
			Metrics			Scores		
Year	Metric	Weight	CN	EU	US	CN	EU	US
2017–18	VC + PE Funding (Billions)	5	$13.5	$2.8	$16.9	2.0	0.4	2.5
2017–18	Number of VC + PE Deals	2	390	660	1,270	0.3	0.6	1.1
2000–19	Number of Acquisitions of AI Firms	2	9	139	526	0.0	0.4	1.6
2017	Number of AI Start-ups	4	383	726	1,393	0.6	1.2	2.2
2019	Number of AI Firms That Have Received More Than $1 Million in Funding	4	224	762	1,727	0.3	1.1	2.5
1960–2018	Highly Cited AI Patent Families	3	691	2,985	28,031	0.1	0.3	2.7
1960–2018	Patent Cooperation Treaty AI Patents	5	1,085	1,074	1,863	1.3	1.3	2.3
	Total Scores	25				4.8	5.3	14.9

Source: European Commission

Graphcore, Jason Lu

- **They have raised $310 million, and have 230+ employees worldwide.**
  - "Today we study static data and deploy a network."
  - "Tomorrow data will be sequenced and computers will learn from experience."
- **IPU, the Colossus GC2. Has 23.6 billion transistors in the processor ("the world’s most complex processor").**
  - Does not support off-chip memory, all memory is on chip.
  - 45 Tb/s memory bandwidth.
  - 125 Pflops at 120 watts.
- **Uses the Poplar software stack, which is similar to CUDA but it is a developer model.**
  - Based on a C++ and python framework.
  - Poplar is an optimized graph mapping software stack.

New EU Processors



E·S·T Common Platform

Key Markets

HPC System PreExascale

Rhea family

ARM(*)

RISC-V

Automotive POC

HPC System Exascale

ARM(*)

RISC-V

Cronos family

Automotive CPU

Gen 1

Gen 2

Gen 3

Few IPs Integration

Some IPs

Many IPs

EPI Common Platform

EPI IP's launch pad

Pan European
Research Platform for HPC & AI

External IPs

2021

2022

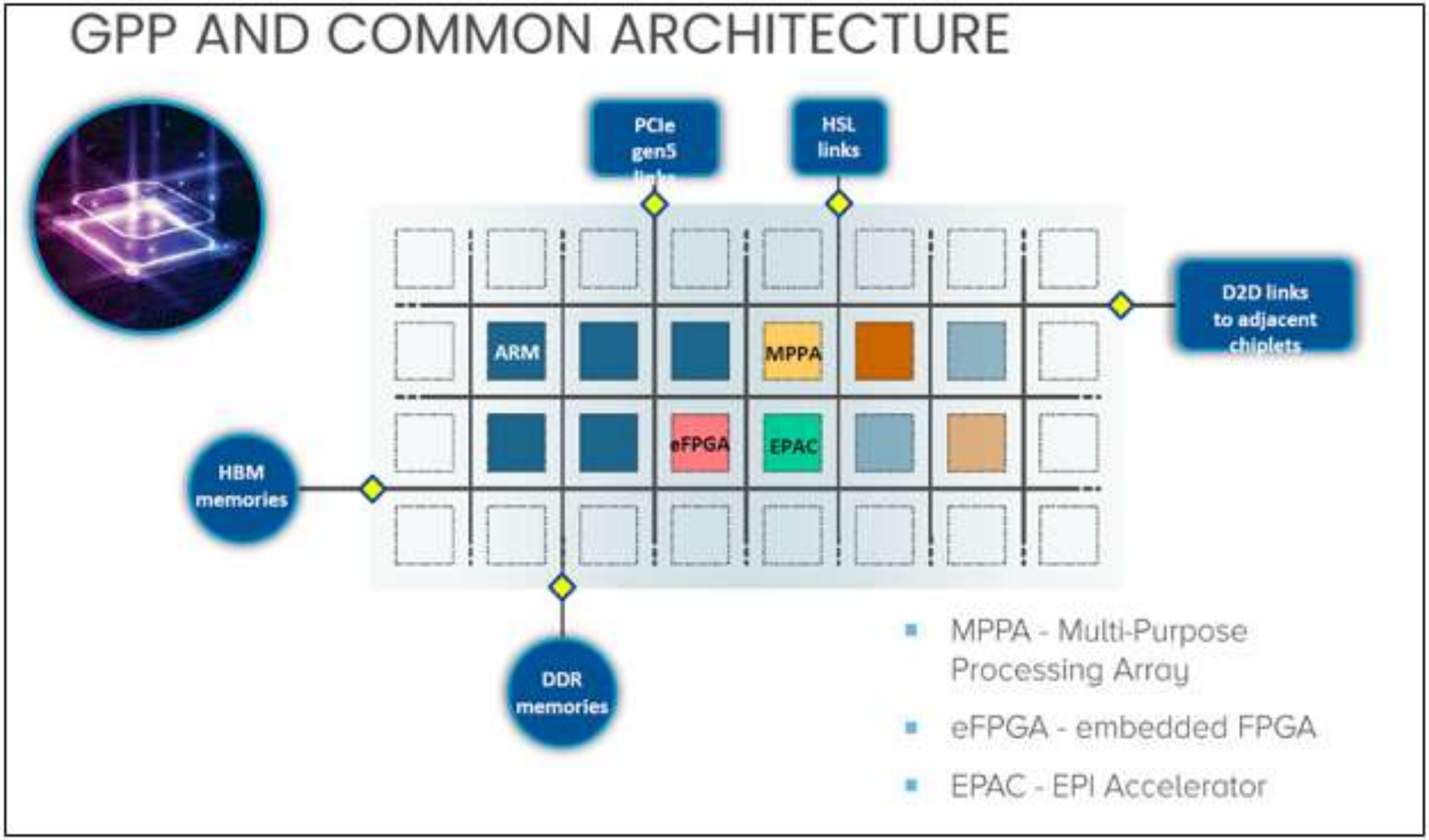
2023

2024

p17| 06-09-2018 | © European Processor Initiative

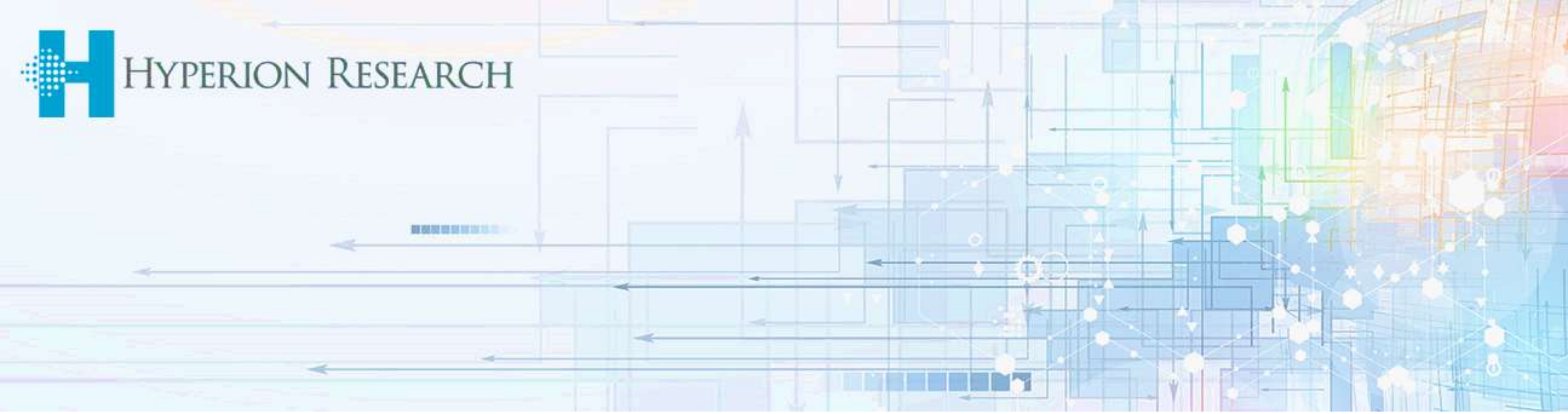
EPI European Processor Initiative

EPI General Purpose Processor (GPP) and Variants



GPP AND COMMON ARCHITECTURE

- MPPA - Multi-Purpose Processing Array
- eFPGA - embedded FPGA
- EPAC - EPI Accelerator



Our New AI Study:
How And Where AI Can Help Advance Science -- Tracking AI Activities Around The World

Focus of the Study

The focus is on where AI can help science

- **Where and how AI technologies can (and do) support DOE mission work**
  - It also includes looking at other types of AI that could help support science in the future
  - And showing which new AI technologies are NOT a good fit for science
- **It will look at developments around the world, both as potential resources for new AI technologies and as potential threats**
  - It will include researching new AI technologies from US and from foreign organizations

Questions to be Researched

- **How will AI change HPC systems?**
- **And how to best construct future AI/HPC systems?**
  - Architectures (data-friendliness, support for concurrent simulation & analytics runs, memory hierarchies)
  - Heterogeneity (workloads, components, precision levels)
  - Processors/coprocessors (CPU, GPU, FPGA, TPU, neuromorphic, ASIC, eASIC)
  - Software (OS, middleware, file systems, automation, integrating orthogonal simulation & analytics results)
- **Mapping AI applications to architectures/technologies**
- **Facility issues, e.g., will sites need multiple system types?**
- **How can HPC decision making be improved with AI technologies?**

Questions to be Researched

- **When will AI get smarter?**
  - Models and algorithms
  - Inferencing
  - Who will actually develop the software and scientific applications?
- **What is the status and future of AI benchmarks?**
  - Who will drive them?
- **Where will AI fit first (and in 5 years & in 10 years)?**
  - What are the best fit & <u>most likely scientific application areas</u>?
  - How does (and will) the US stack-up?
  - Who are the major foreign competitors and where do they stand?

Questions to be Researched

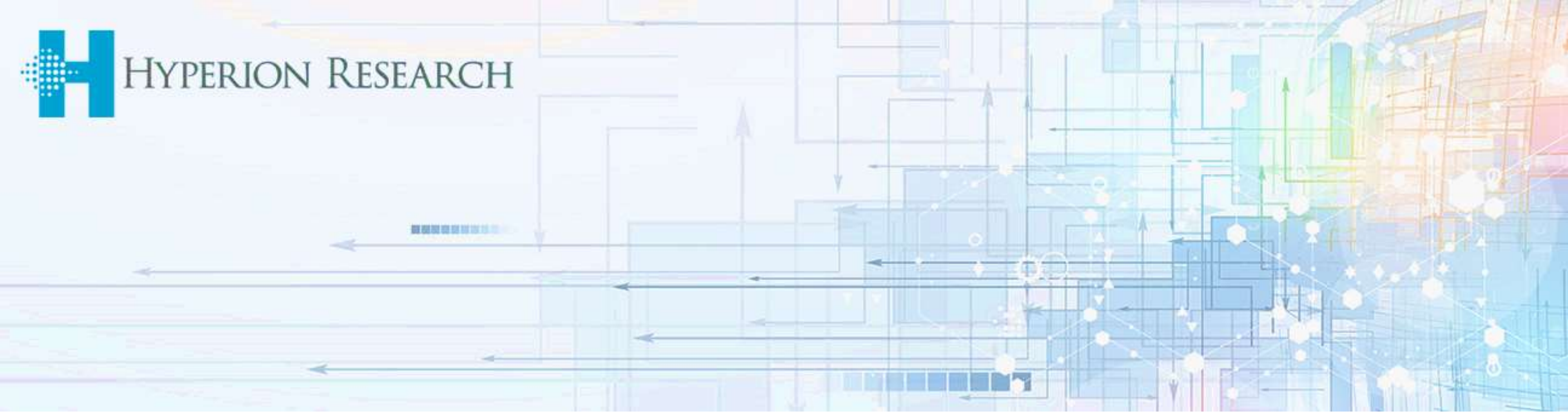
- **How will verification, validation and certification be accomplished?**
  - Including uncertainty quantification
  - Will it require a side-by-side computer?
  - How will legal and regulatory systems catch up?
  - How to address explainability?
  - Where is bias, and what can be done about it?
- **Will AI, ML & deep learning keep growing very fast, or will transparency, uncertainty quantification, and other issues hold them back?**
  - And what can be done in advance to keep these issues from holding AI back?
- **How can the lack of large enough data sets be addressed?**

Questions to be Researched

- **How will supercomputers <u>evolve</u> over the next 2 to 5 years to handle AI and simulation?**
  - How will new AI-focused technologies fit into computers for science?
  - How fast will other AI methods beyond ML & DL (e.g., graphing, semantic analysis) grow?
- **Will AI systems be constructed from a large mix of components, coming from all around the world?**
  - Which components will be of highest value: processors? memories? software? Or something else?
  - How will indigenous technology initiatives affect AI?
  - Will large volume (non-HPC) devices drive core AI technologies?
  - To what extent will HPC and commercial hyperscale architectures converge?

Questions to be Researched

- **Who will drive AI progress: HPC users vs. social media/Internet/cloud companies?**
  - Which domains will have enough data for accurate DL?
  - Are there ways to reduce the needed data size?
  - Convergence: Google, AWS, FB, et al. are adopting HPC as HPC attempts to move into HPDA-AI markets
  - CSP competition: China vs. the world
  - How can DOE leverage these technologies?
- **Who will be the major AI OEMs?**
  - How will existing computer vendors do against new providers?
  - Are Chinese providers a major threat? And Europe, Japan & Russia?



In Summary:
Some Predictions For the Next Year Or So

The Exascale Race
Will Drive New Technologies

- **The global ES race is boosting funding for the Supercomputers market segment and creating widespread interest in HPC**
- **Exascale systems are being designed for HPC, AI, HPDA, etc.**
  - This will drive new processor types, new memories, new system designs, new software, etc.
- **In some cases HPC is too strategic to depend on foreign sources**
  - This has led to indigenous technology initiatives

Storage Systems Will Increasingly Become More Critical

- **Data-intensive HPC is driving new storage requirements**
  - Iterative methods will expand the size of data volumes needing to be stored
- **Future architectures will allow computing and storage to happen more pervasively on the HPC infrastructure**
  - Metadata management will deal with data stored in multiple geographic locations and environments
- **Physically distributed, globally shared memory will become more important**
- **More intelligence will need to be built into storage software**

Artificial Intelligence Will Grow Faster Than Other IT Sectors

- **The AI market is at an early stage but already highly useful (e.g., visual and voice recognition)**
  - Once better understood, there are many high value use cases that will drive adoption
- **Advances in inferencing will reduce the amount of training needed for today's AI tasks**
  - But the need for training will grow to support more challenging tasks
- **The trust (transparency) issue that strongly affects AI today will be overcome in time**
- **Learning models (ML, DL) have garnered most of the AI attention, but graph analytics will also play a crucial role with its unique ability to handle temporal and spatial relationships**

Questions?

ejoseph@hyperionres.com

anorton@hyperionres.com

sconway@hyperionres.com

bsorensen@hyperionres.com

Hyperion Definitions:

AI, Machine Learning, Deep Learning

- **Artificial Intelligence (AI):** a broad, general term for the ability of computers to do things human thinking does (but NOT to think in the same way humans think). AI includes machine learning, deep learning (a.k.a. cognitive computing) and more minor methodologies.
- **Machine learning (ML)**: a process where examples are used to train computers to recognize specified patterns, such as human blue eyes or numerical patterns indicating fraud. The computers are unable to learn beyond their training and human oversight is needed in the recognition process. The computer follows the base rules given to it.
- **Deep Learning (DL):** an advanced form of machine learning that uses digital neural networks to enable a computer to go beyond its training and learn on its own, without explicit programming or human oversight. The computer develops its own rules.

Examples of Recent Hyperion Research Worldwide Studies for U.S. Federal Agencies

- **The Evolution of AI Hardware and Software Ecosystems**
- **The Evolution of Field Competencies in Machine/Deep Learning and Resultant Industries**
- **AI Primer for Senior Military Decision-Makers**
- **AI Hardware Technology, Vendor Status and Trends**

Cloud Companies Joining the Processor Development Party

- **Google developed tensor cores to accelerate machine learning workloads.**
  - Only available on Google cloud for now
  - Google announced the third generation TPU last year.
- **Amazon, at their re:Invent conference in November of 2018, announced their inference chip, Inferentia.**
  - Designed to accelerate machine learning, especially inferencing.