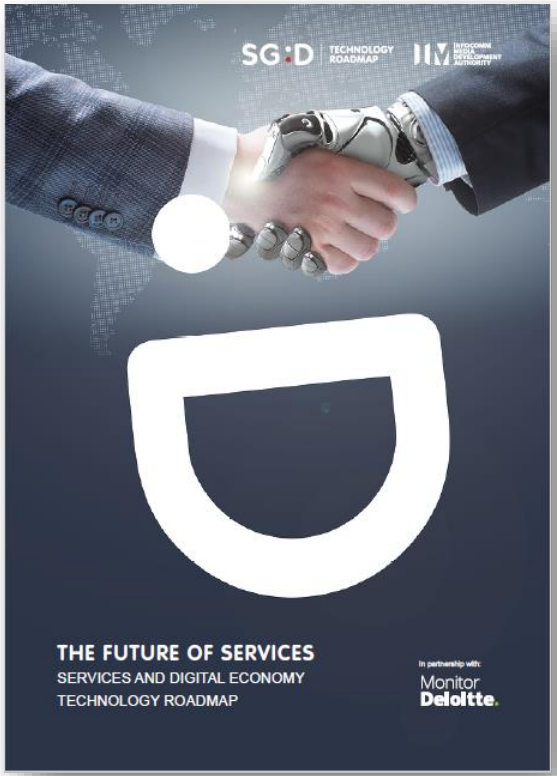
HETEROGENEOUS COMPUTE INFRASTRUCTURE FOR SINGAPORE

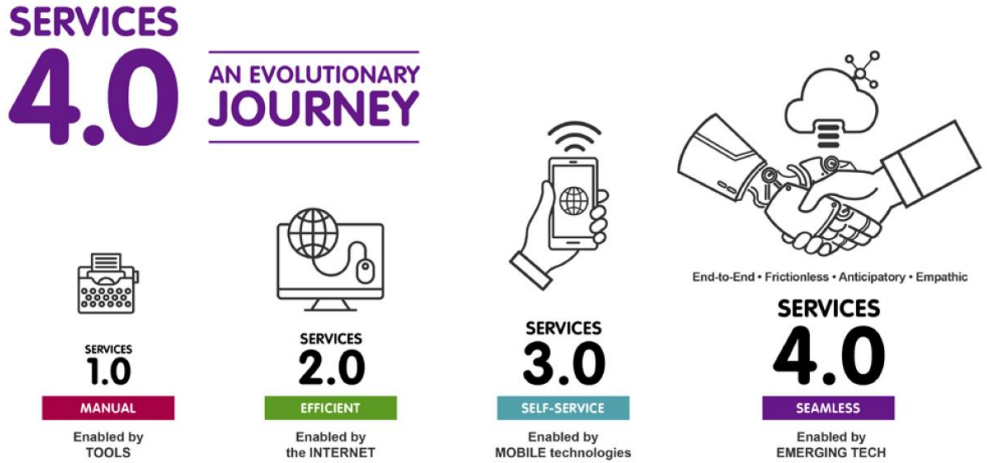
PHILIP HEAH
ASSISTANT CHIEF EXECUTIVE
TECHNOLOGY & INFRASTRUCTURE GROUP

SG:D
EMPOWERING POSSIBILITIES

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

LAUNCH OF SERVICES AND DIGITAL ECONOMY (SDE) TECHNOLOGY ROADMAP (NOV 2018)





Source – IMDA Services and Digital Economy (SDE) Technology Roadmap 2018

SG:D EMPOWERING POSSIBILITIES

IMDA INFOCOMM MEDIA DEVELOPMENT AUTHORITY

VISION: SINGAPORE AS A SERVICES 4.0 HUB

A Launchpad for Services 4.0

A **#Service40Hub** where **#EveryBusinesADigitalBusiness** and **#EmpoweringPossibilities** for Businesses

A Competitive Workforce Augmented with Technology

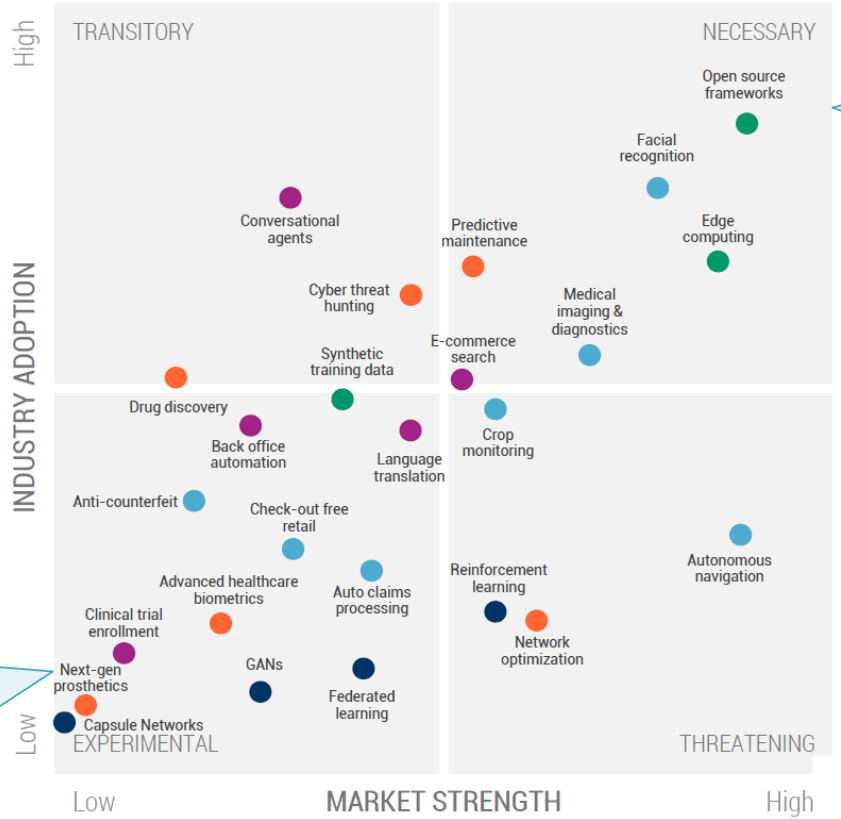
A **#DigitalTalentHub** where there is a **#BotForEveryWorker** and **#EmpoweringPossibilities** for Workers

A Vibrant ICM Ecosystem where Emerging Tech is made easily Accessible

#EmpoweringPossibilities with **#GoCloudNative**

Source – IMDA Services and Digital Economy (SDE) Technology Roadmap 2018

AI EXPECTED TO BE ONE OF THE KEY TECHNOLOGY ENABLERS FOR SERVICES 4.0



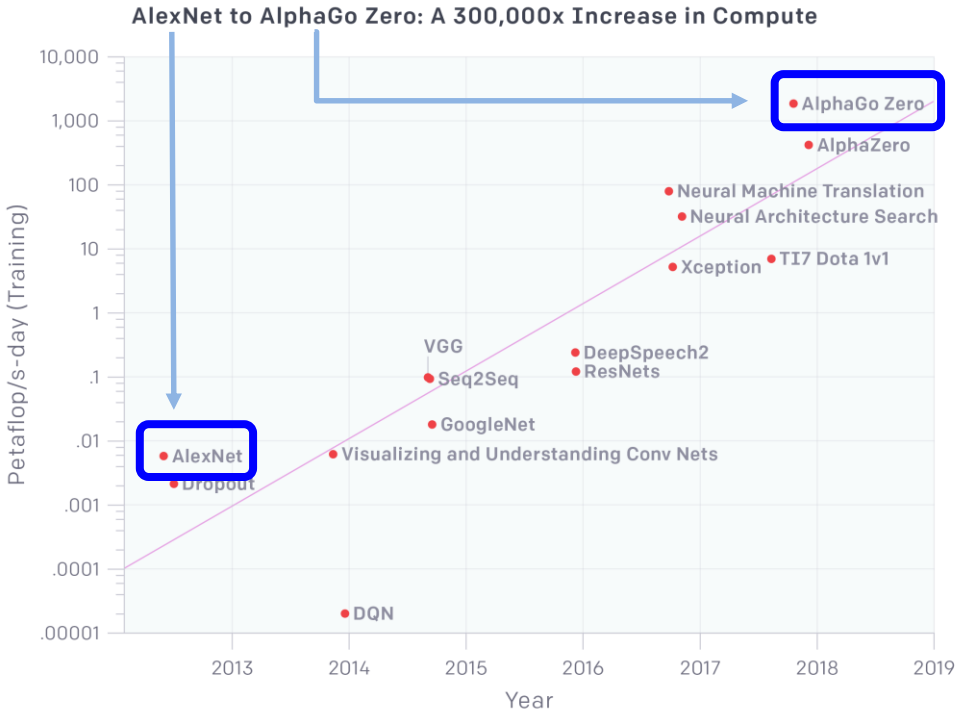
Trends with widespread industry and customer implementation and adoption

Conceptual or early-stage trends already spurring early interest and POCs

- Application: Computer Vision
- Application: National Language Processing/Synthesis
- Application: Predictive Intelligence
- Architecture
- Infrastructure

Source – "AI Trends 2019", CB Insights, 2019

AMOUNT OF COMPUTE USED IN LARGEST AI TRAINING RUNS HAS BEEN INCREASING EXPONENTIALLY



Reference – "AI and Compute", OpenAI, May 2018

Open AI Analysis

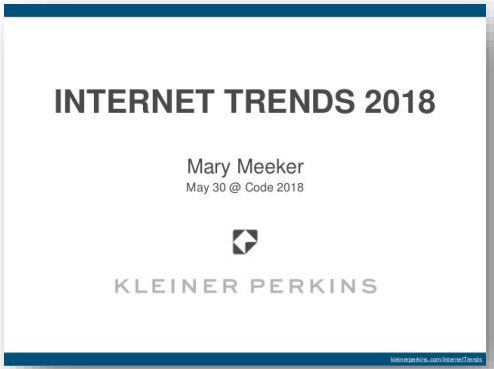
- Since 2012, the amount of compute used in the largest AI training runs has been **increasing exponentially**
- Doubles every 3.5 months
  - Moore's Law: doubles very 18 months
- **Improvements in compute have been a key component of AI progress**

SG:D EMPOWERING POSSIBILITIES

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

NEED TO LAY FOUNDATION FOR SG'S DIGITAL INFRASTRUCTURE TO SUPPORT SERVICES 4.0 IN DIGITAL ECONOMY

COMPUTE POWER IS ONE OF THE KEY ECONOMIC GROWTH DRIVERS IN THE 21ST CENTURY



CENTURY	ECONOMIC GROWTH DRIVERS
Pre-18th	Cultivation & Extraction
18th - 19th	Manufacturing & Industry
21st …	**Compute Power & Human Potential**

Source – Internet Trends 2018, Mary Meeker

SG:D EMPOWERING POSSIBILITIES

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

IMPLICATIONS OF COMPUTE POWER

01

SCIENTIFIC ADVANCEMENT

Computational leadership is critical to scientific advancement and leadership

02

INDUSTRIAL COMPETITIVENESS

Enable organisations to lower cost of innovation and develop high value innovation

03

NATIONAL SECURITY

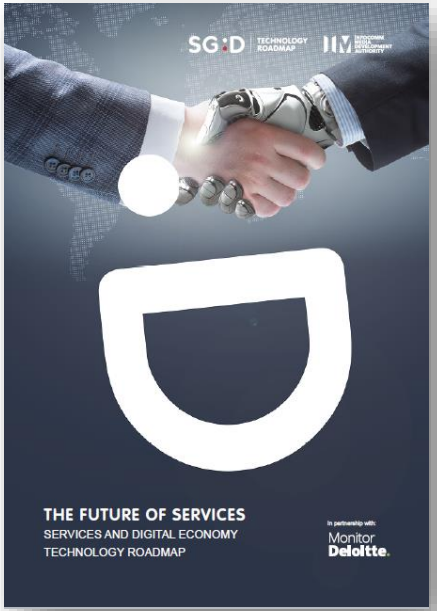
Cyberspace now the 5th battleground in addition to land, sea, air and space

Reference – "The Exascale Computing Race Isn't About Bragging Rights", DZone, Feb 2018

SG:D EMPOWERING POSSIBILITIES

IMDA INFOCOMM MEDIA DEVELOPMENT AUTHORITY

SDE TECHNOLODY ROADMAP RECOMMENDATION – NEED FOR SINGAPORE TO DEVELOP COMPUTE INFRASTRUCTURE



INVEST TO INCREASE CAPACITY AND CAPABILITY OF DATA CENTRES AND HYPERSCALE COMPUTING

Cloud Native Architecture will **demand increased capacity and computing power**. This, in turn, will create a pressing need for more sophisticated data centres.

Initiatives such as high rise and floating data centres could mitigate Singapore's land scarcity and hot climate.

Data centres built on **hyperscale computing architecture** would be another area of focus to meet the surge in demand.

Source – "The Future of Services", IMDA, Dec 2018

SG:D EMPOWERING POSSIBILITIES

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

PROPOSED APPROACH FOR SINGAPORE – HETEROGENEOUS COMPUTE INFRASTRUCTURE

SG:D
EMPOWERING POSSIBILITIES

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

WHY HETEROGENOUS COMPUTE INFRASTRUCTURE?

SHIFT TOWARDS SPECIALIZED COMPUTE TO FUEL DEMANDS OF AI

SPECIAL PURPOSE CLOUD	AI CHIPS	EDGE COMPUTE	ARCHITECTURES	COMPLEXITY IN SOFTWARE
• Google TPU • Microsoft Brainwave • Intel Nervana • IBM Power AI • Nvidia v100	• CPU • GPU • FPGA • Custom ASICs	• Hardware accelerators • AI SOC	• Cluster compute • HPC • Neuromorphic • Quantum compute	• Model tuning/optimisations specific to hardware • Growing need for compilers to optimize based on deployment hardware • Workload specific compute: Model training, inference

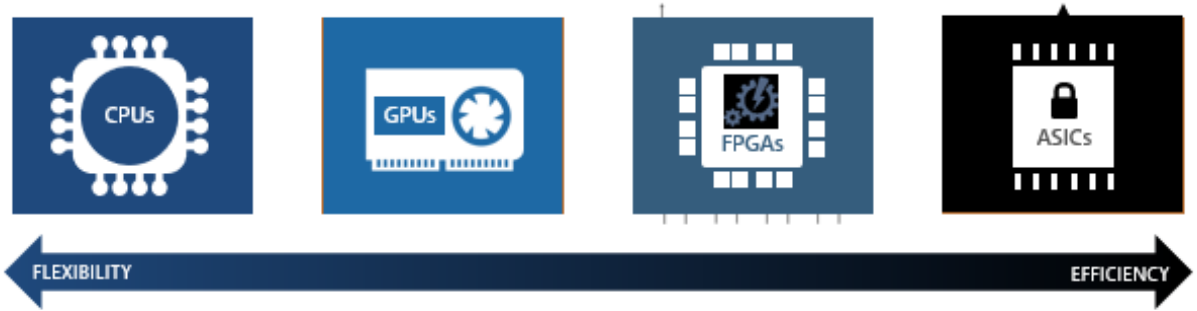
Reference – "Distributed Deep Learning Optimisations for Finance", Geeta Chauhan, CTO SVSG

SG:D EMPOWERING POSSIBILITIES

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

SHIFT TOWARDS SPECIALISED COMPUTE FOR AI

EXISTING PROCESSORS NOT ORIGINALLY DESIGNED FOR NEW AI APPLICATIONS



	CPUs	GPUs	FPGAs	ASICs
+	• General-purpose processors • Sufficient for inference	• Highly parallel, high performance	• Reconfigurable • Good for constantly evolving workloads • Efficient	• Best performance • Most energy and cost efficient • Fully customisable
−	• Serial processing less efficient than parallel processing	• Less efficient than FPGAs	• Difficult to program • No major AI framework	• Long development cycle • Requires high volume to be practical
TRAINING RANK	1	2	3	
INFERENCE RANK	3	2	1	

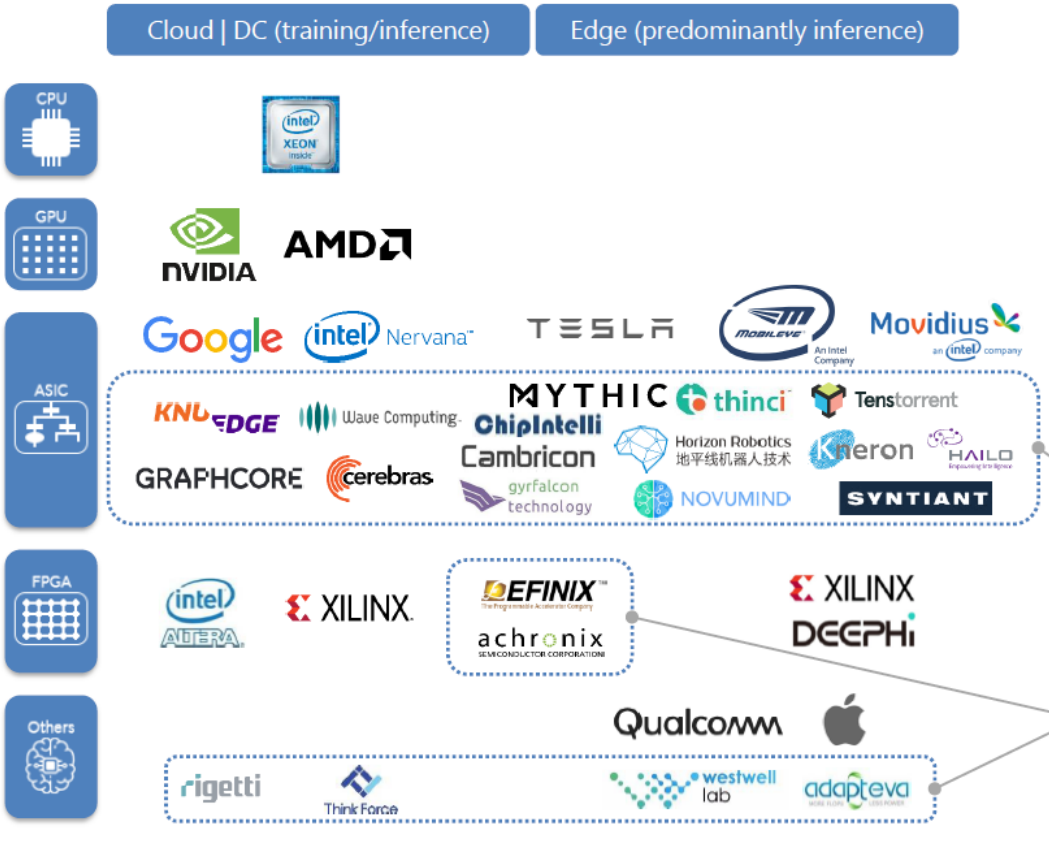
Sources –
"What is FPGA and Project Brainwave?", Microsoft, Sep 2018
"AI-Optimised Chipsets, Part I", Vertex Ventures, Mar 2018

SG:D EMPOWERING POSSIBILITIES

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

SHIFT TOWARDS SPECIALISED COMPUTE FOR AI

PLAYERS DEVELOPING AI-OPTIMIZED HARDWARE



Key Observations

- At least **45 startups** are working on chipsets purpose-built for AI tasks
- At least 5 of them have raised more than USD 100M from investors
- According to CB Insights, VCs invested more than **USD 1.5B** in chipset startups in 2017, nearly doubling the investments made 2 years ago

Most startups seem to be focusing on ASIC chipsets at the edge and in the cloud/DC

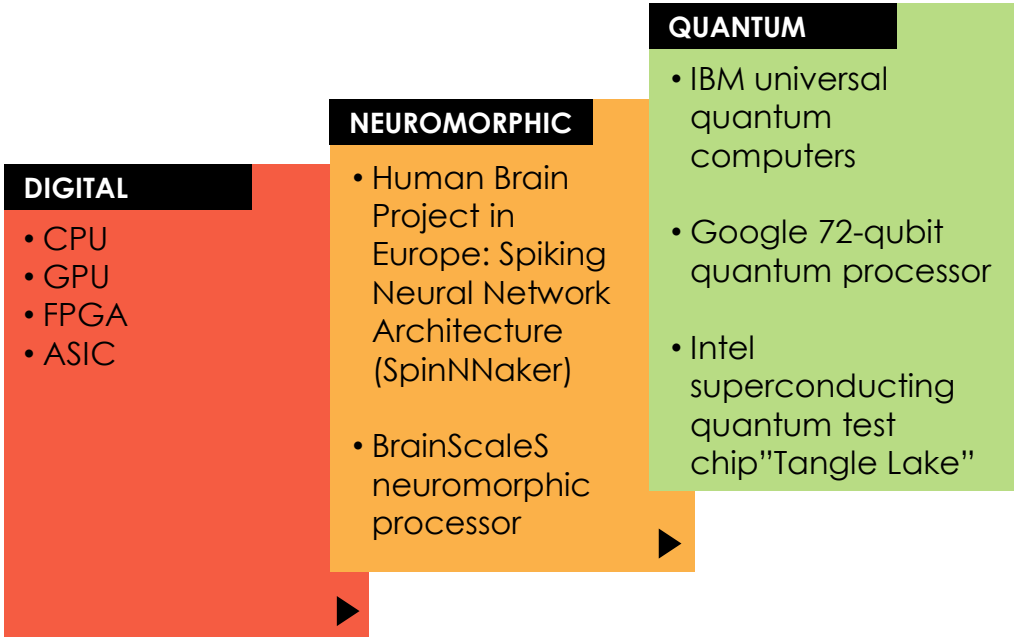
FPGAs and other architectures also appear attractive to chipset startups

Sources – "AI-Optimised Chipsets, Part III", Vertex Ventures, Aug 2018

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

SHIFT TOWARDS SPECIALISED COMPUTE FOR AI

EMERGING ARCHITECTURES TO ENABLE AI



"Three-way technology race to bring faster, easier, cheaper, and smarter AI"

- High Performance Computing is available today but so are new commercial versions of actual Quantum computers and Neuromorphic Spiking Neural Nets
- These two new entrants are going to revolutionize AI and deep learning starting now

- Data Science Central article

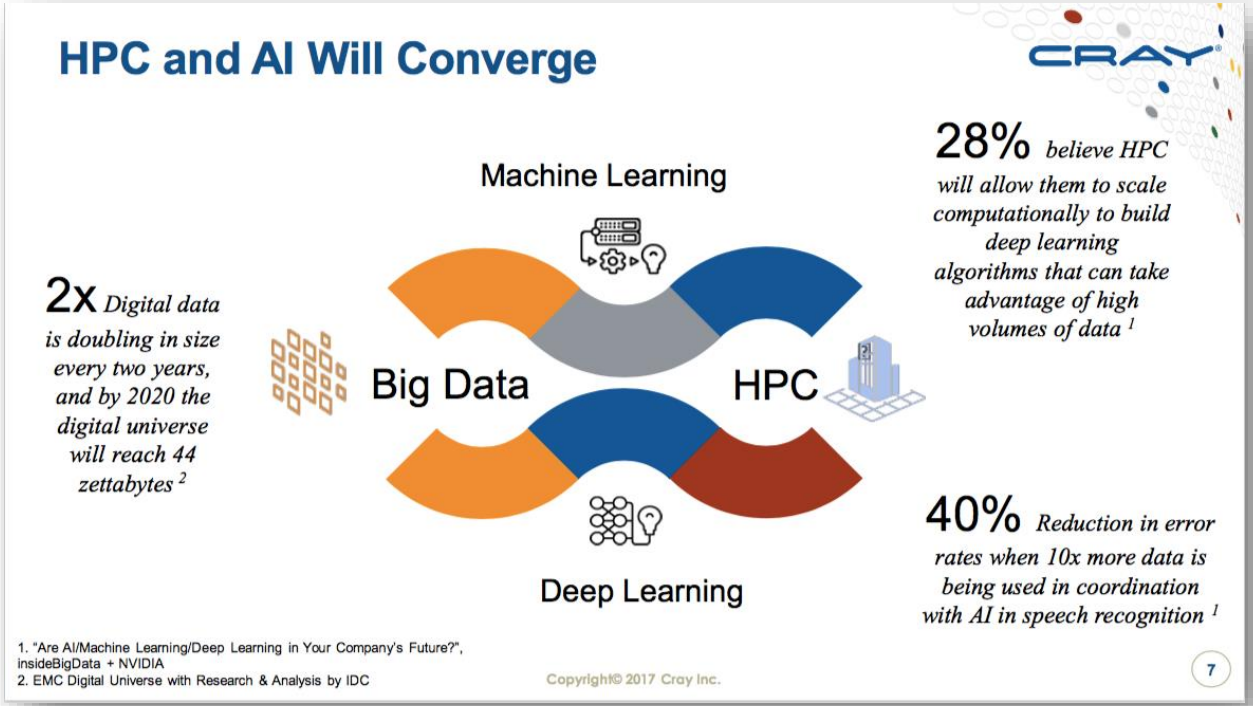
Source – "The Three Way Race to the Future of AI: Quantum vs Neuromorphic vs High Performance Computing", DataScienceCentral, Nov 2017

SG:D EMPOWERING POSSIBILITIES

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

CONVERGENCE OF HPC & AI EXPECTED TO DRIVE HPC ADOPTION BY COMMERCIAL PLAYERS

HPC for Industrial and Commercial Applications



- Driven by convergence of HPC and AI
- Performance expected to be an AI innovation and adoption driver
- Deep learning enabled by HPC

Source – Perspective on HPC-Enabled AI, Tim Barr, Sep 2017

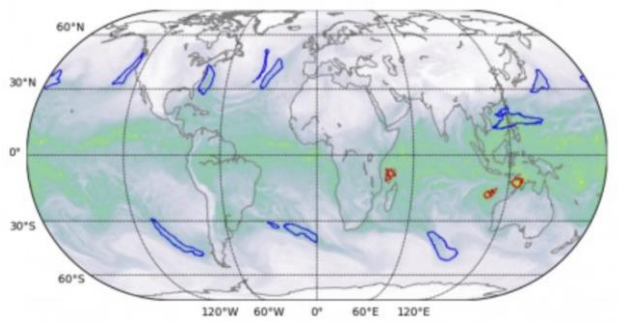
SG:D EMPOWERING POSSIBILITIES

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

HPC/AI CONVERGENCE – ENABLE TECH ADVANCEMENT

TOM SIMONITE BUSINESS 01.31.19 05:58 PM

THE WORLD'S FASTEST SUPERCOMPUTER BREAKS AN AI RECORD



High-quality segmentation results produced by deep learning on climate datasets.

- Summit (Oak Ridge National Lab) powered deep learning experiment to detect weather patterns like cyclones output from climate simulations at record-breaking speeds

- Demonstrated
  - Scientific **potential of adapting deep learning to supercomputers**
  - **Machine learning can benefit from more computing power**

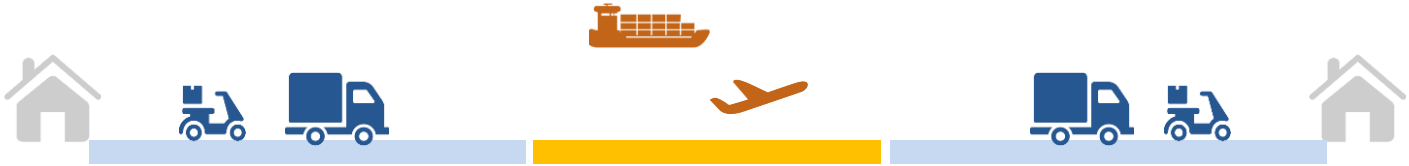
Source – "World's Fastest Supercomputer Breaks AI Record", Wired, Jan 2019

SG:D EMPOWERING POSSIBILITIES

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

NEED FOR HETEROGENEOUS HYPERSCALE COMPUTE INFRASTRUCTURE TO ADDRESS DIVERSE NEEDS OF SERVICES 4.0 IN DIGITAL ECONOMY

LOGISTICS
MULTI-MODAL
TRANSPORT



Trade offs: Item(s) delivered, time to delivery, cost of delivery, etc

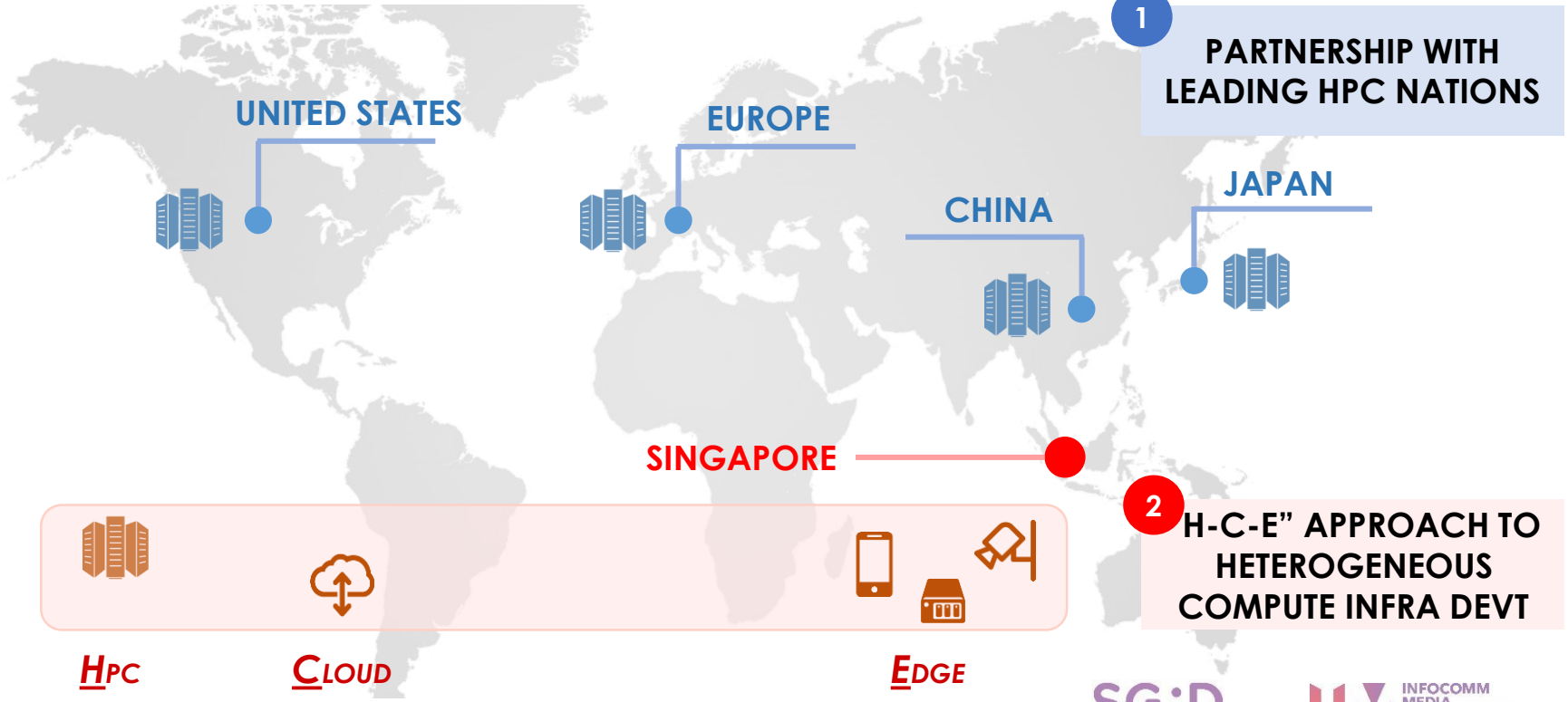
COMPUTE INFRASTRUCTURE
"MULTI-MODAL"
HETEROGENEOUS

SG:D EMPOWERING POSSIBILITIES

INFOCOMM MEDIA DEVELOPMENT AUTHORITY

APPROACH

HETEROGENEOUS HYPERSCALE COMPUTING FOR SINGAPORE



1 PARTNERSHIP WITH LEADING HPC NATIONS

2 H-C-E" APPROACH TO HETEROGENEOUS COMPUTE INFRA DEVT

SG:D EMPOWERING POSSIBILITIES

INFOCOMM MEDIA DEVELOPMENT AUTHORITY


THANK YOU
sgdigital.sg
twitter.com/IMDAsg
facebook.com/IMDAsg
youtube.com/IMDAsg
SG:D
EMPOWERING POSSIBILITIES
INFOCOMM
MEDIA
DEVELOPMENT
AUTHORITY