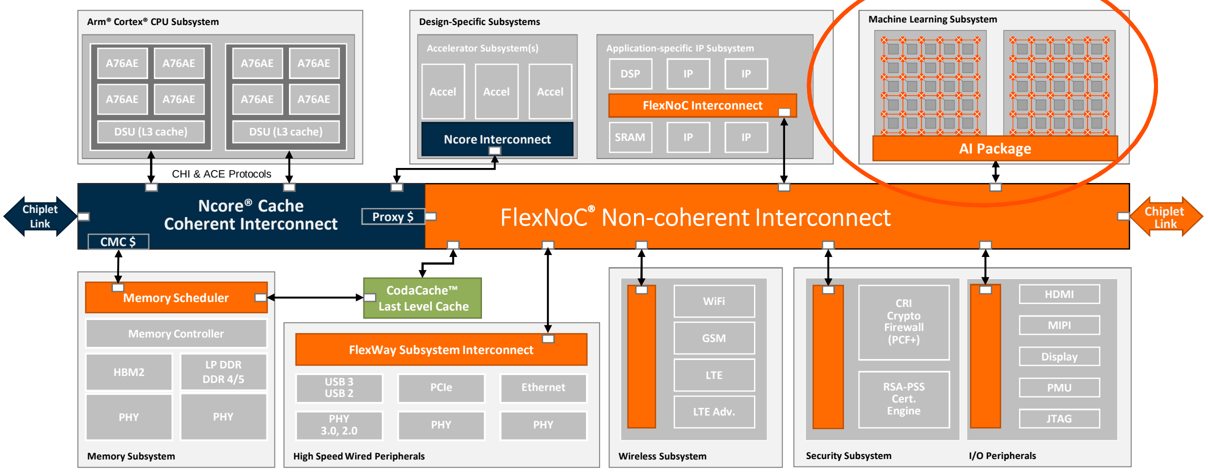
Interconnect for Machine Learning

ML = INTERCONNECT + NEURONS + SOFTWARE

K. CHARLES JANAC

President and CEO

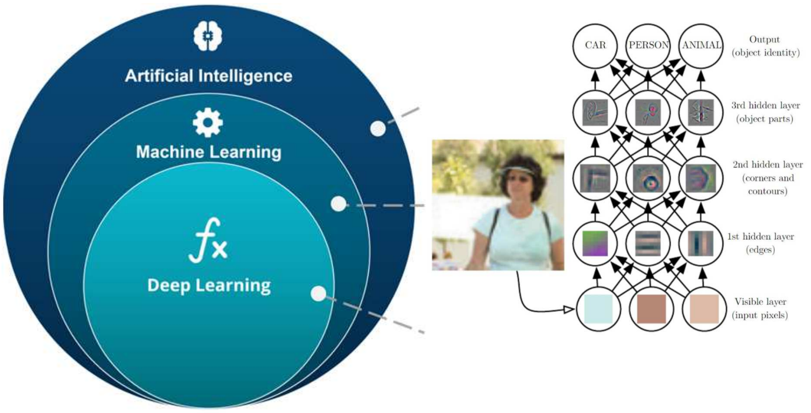
Arteris IP, the Data Highway of the SoC

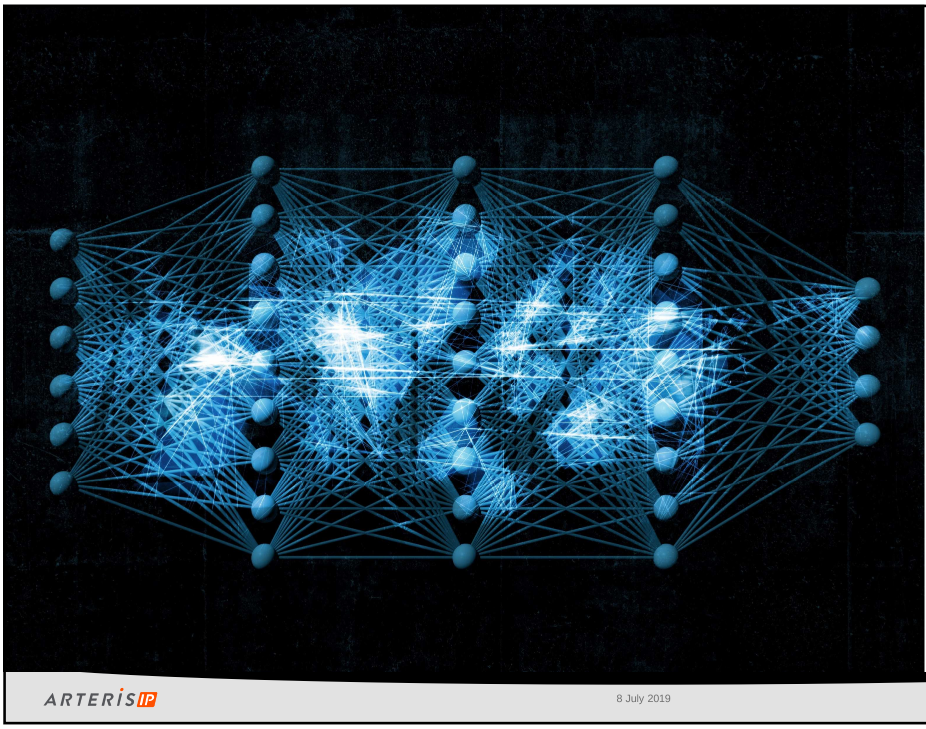


- ■ Arteris IP Ncore® cache coherent interconnect IP
- 🟧 Arteris IP FlexNoC® non-coherent interconnect IP
- 🟩 Arteris IP CodaCache™ last level cache

Machine Learning – Training vs Inference

- Deep Learning is a branch of machine learning that involves layering algorithms in an effort to gain greater understanding of the data.
  - Deep learning can take raw information without any meaning and construct hierarchical representations that allow insights to be generated
  - Deep learning's results improve more and more with large training data sets
  - Machine learning allows machines to make predictions based on "experience" data sets





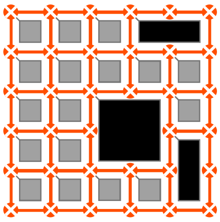
The Brain is neurons and interconnect

Machine Learning SoC is

- Neurons (processors)
- Interconnect
- I/Os and peripherals
- Lots of software
- Lots of data

Deep Learning Specific Interconnect Features

FOR ACCELERATED DEVELOPMENT OF MACHINE LEARNING SOCS



Regular (AI) Topologies

- NEW! Automated Topology generation
- NEW! Customization of automated results
- NEW! Flexible router architecture

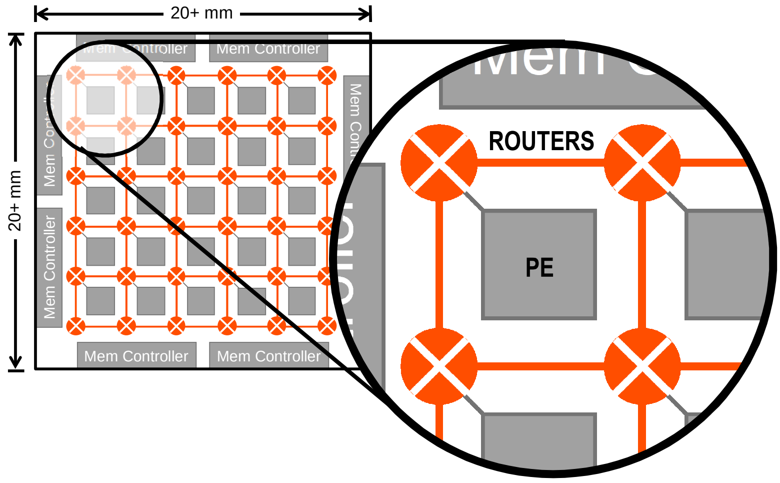
Huge Bandwidth

- NEW! Multicast
- NEW! Multi-channel HBM2 memory support
- NEW! High bandwidth datapaths

Large Chips

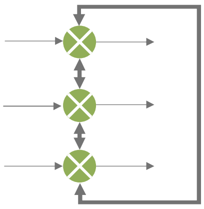
- NEW! Source synchronous communications
- NEW! VC-Links™ - Virtual Channels

Characteristics of mesh and ring architectures

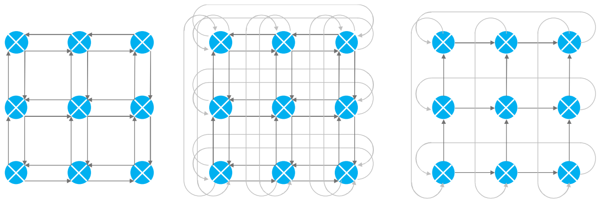


- One or more processing elements (PE), memories (caches) or I/O controllers per corner router
- Multicast / Broadcast writes for BW efficiency
- Sophisticated interleaving for optimal off-chip HBM2 access

Regular Topology Generation



- Simple elementary components enable building essentially any topology provided that:
  - Routing function is static and defined at configuration time
  - Routes have no dependencies – deadlock avoidance is guaranteed by the tool
- It is then possible to build regular topologies such as:
  - Large crossbars with distributed pipelining
  - 2D Meshes
  - Rings (1D) and Torus (2D)

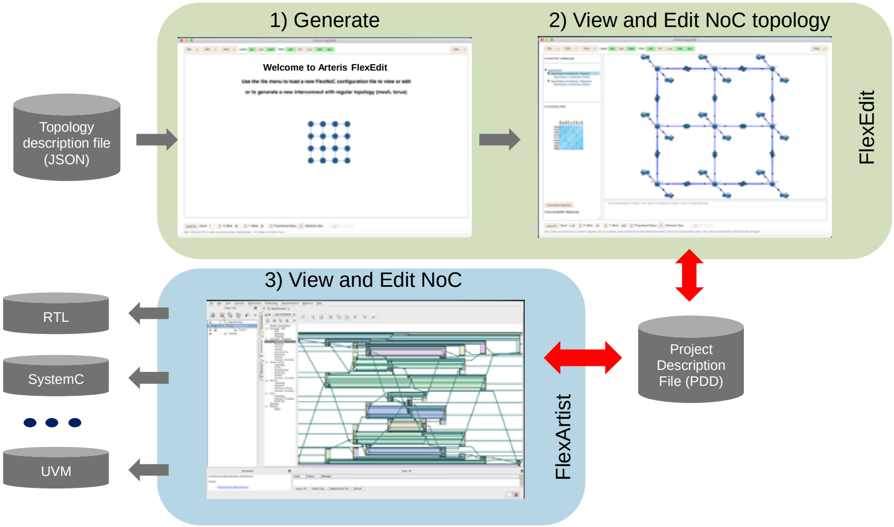


a) Mesh topology

b) Bi-directional Torus topology

c) Torus topology

Editing Regular Topology Generator Results



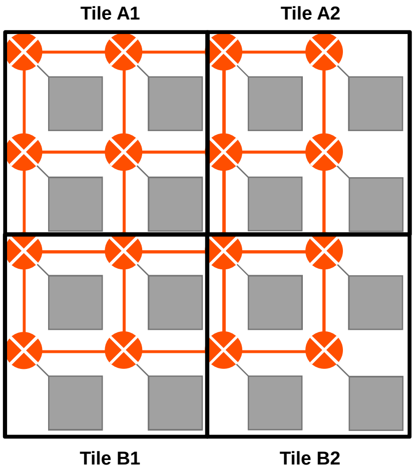
1) Generate
2) View and Edit NoC topology
FlexEdit
Topology description file (JSON)
3) View and Edit NoC
FlexArtist
RTL
SystemC
UVM
Project Description File (PDD)

Welcome to Arteris FlexEdit

Use the file menu to load a new FlexNoC configuration file to view or edit

or to generate a new interconnect with regular topology (mesh, torus)

Mesh tiling benefits and drawbacks



PRO

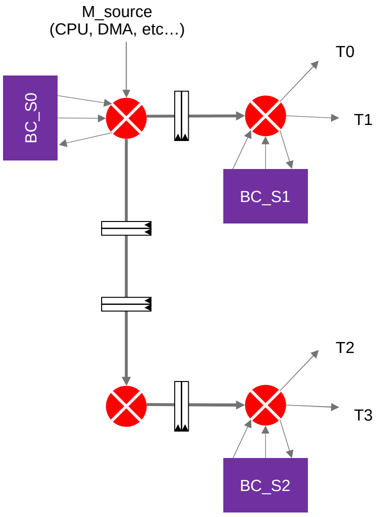
- Can be easier for place & route team to massively scale (hard macros)
- Can simplify top-level interconnect, or allow scalable reuse of existing NoC architecture

CONS

- Requires additional logic and memory in tile
- Adds complexity to NoC addressing
- Reduces flexibility for system-level optimizations (QoS, memory interleaving, power management, etc.)

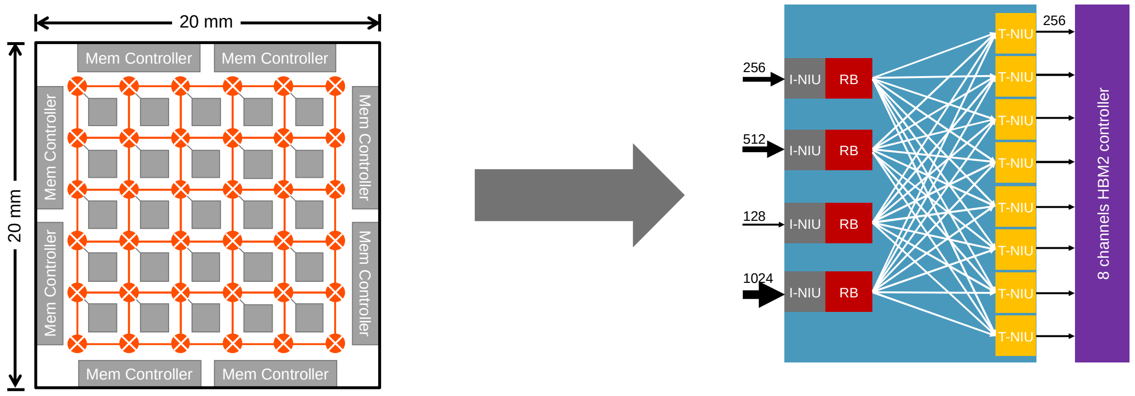
Write Broadcast Station for Multicast

- Broadcast Station **ingress ports** mapped as Write Only slave in FlexNoC memory map
  - There can be any number of stations in a FlexNoC
  - Broadcast done **as close as possible to the destinations**
- Writing to Broadcast Station will make it send in turn the writes to multiple destinations
  - Based on address prefix per egress port
  - Can be other broadcast stations
  - Can be "normal" slaves
- Broadcast Stations **egress ports** as masters of the NoC
  - Optional: selection of egress ports by user bits: dynamic multicast
- Optional : Support of posted writes for higher performance



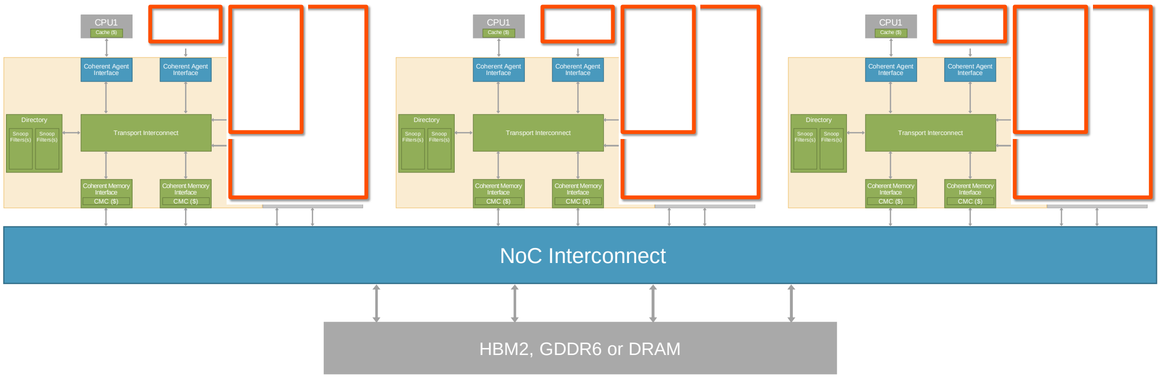
Getting data off-chip: HBM2 load balancing

From this… **…to this.**

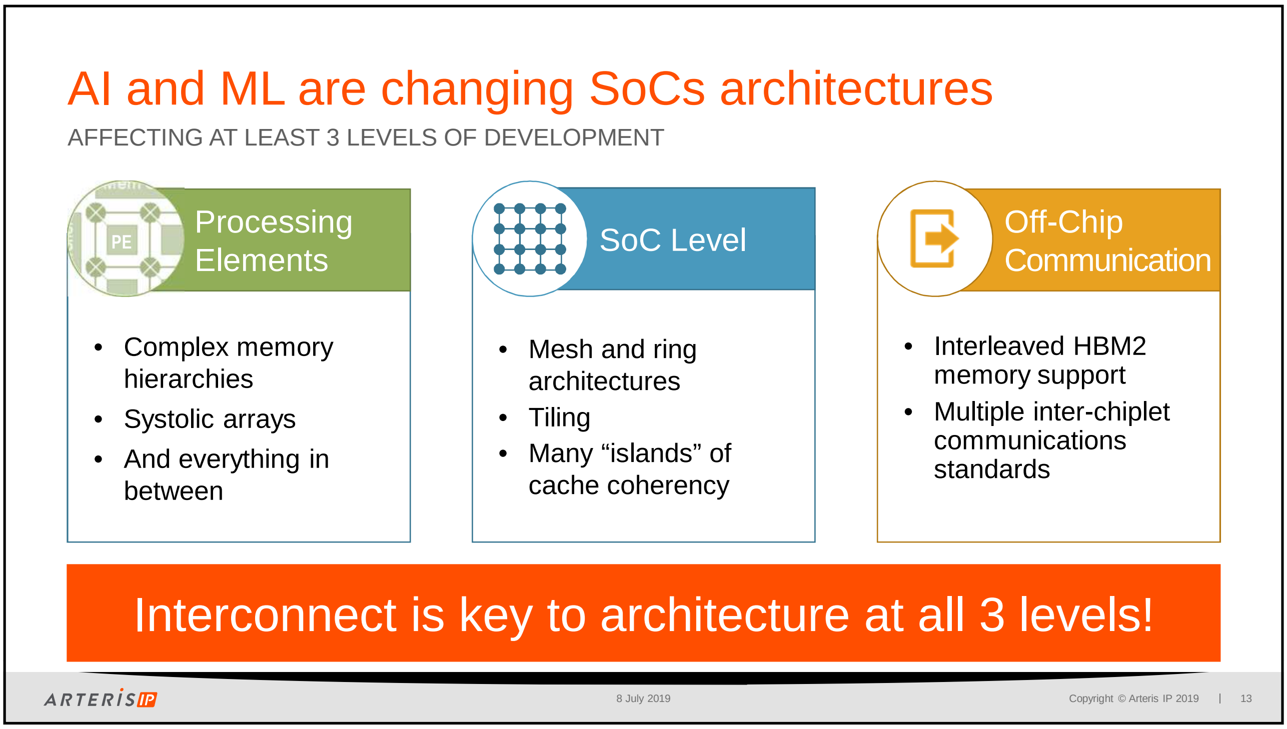


- Advanced interleaving and response reordering
- Distributed controller ports → long distances → source-synchronous and virtual channels

Scale-up alternative: “Islands” of cache coherency



- Multiple subsystems that are internally cache-coherent
- Each subsystem usually different, optimized for set(s) of tasks
- Can provide scalability for edge inference processing while meeting latency and power requirements



AI and ML are changing SoCs architectures

AFFECTING AT LEAST 3 LEVELS OF DEVELOPMENT

Processing Elements

- Complex memory hierarchies
- Systolic arrays
- And everything in between

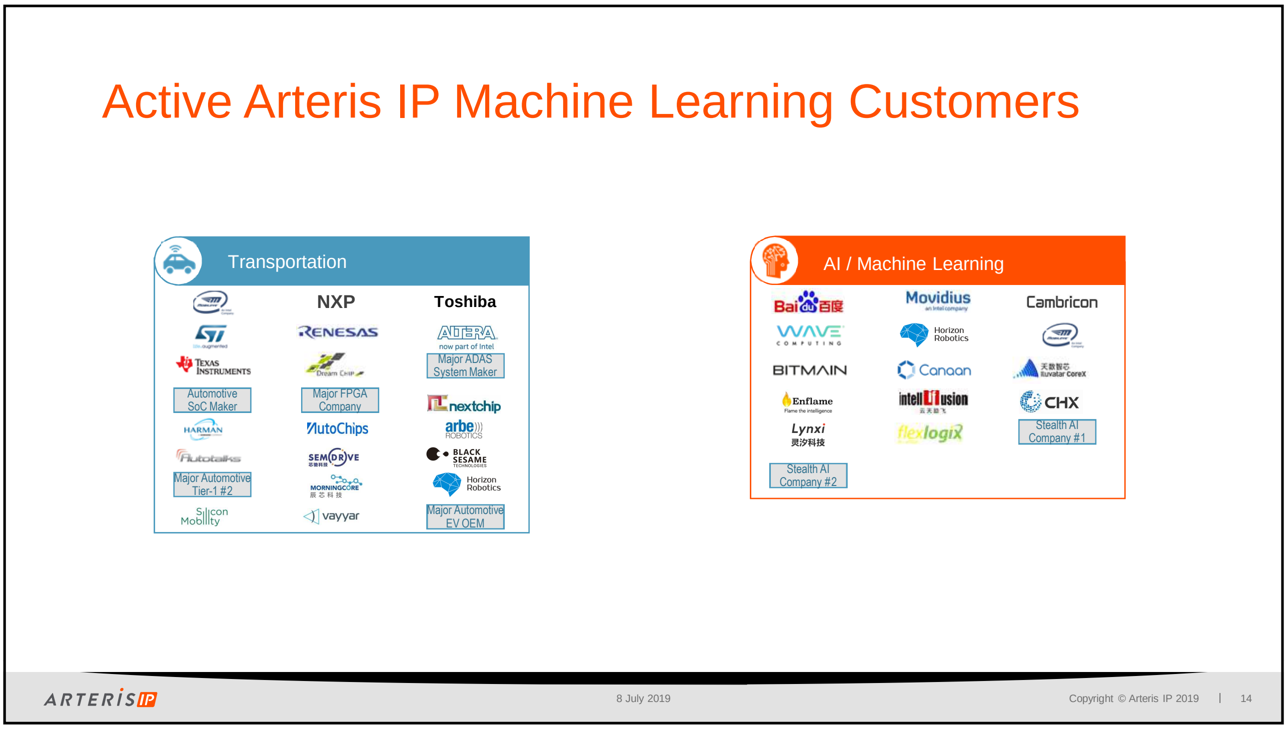
SoC Level

- Mesh and ring architectures
- Tiling
- Many “islands” of cache coherency

Off-Chip Communication

- Interleaved HBM2 memory support
- Multiple inter-chiplet communications standards

Interconnect is key to architecture at all 3 levels!



Active Arteris IP Machine Learning Customers

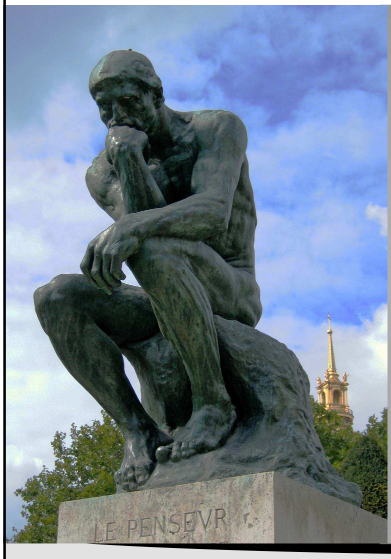
Transportation

Mobileye An Intel Company	NXP	Toshiba
ST life.augmented	Renesas	Altera now part of Intel
Texas Instruments	Dream Chip	Major ADAS System Maker
Automotive SoC Maker	Major FPGA Company	nextchip
Harman	AutoChips	arbe ROBOTICS
Autotalks	SemiDrive 芯驰科技	Black Sesame Technologies
Major Automotive Tier-1 #2	MorningCore 辰芯科技	Horizon Robotics
Silicon Mobility	vayyar	Major Automotive EV OEM

AI / Machine Learning

Baidu 百度	Movidius an Intel company	Cambricon
Wave Computing	Horizon Robotics	Mobileye An Intel Company
BITMAIN	Canaan	天数智芯 Iluvatar CoreX
Enflame Flame the intelligence	intellifusion 云天励飞	CHX
Lynxi 灵汐科技	flexlogix	Stealth AI Company #1
Stealth AI Company #2		

Challenges: Determinism, Visibility, Controllability



- ***How do you verify a deep learning system?***
- ***How do you debug the Neural Network?***
- ***How to you trace intermediate results?***
- ***How do you determine how ML arrives @ solution?***
- ***How do you make a Neural Network “safe”?***
- ***What are the ethics and biases of these systems?***

Conclusion

- ML today is not particularly intelligent, but it is clever and improving
  - CNNs being deployed in automotive and data center acceleration SoCs
  - Eventually majority of complex SoCs will contain machine learning sections
  - A simple continuation of the exponential increasing processing performance trend will not break ML out of the scope of applications that it can do well today
  - CNNs are non-linear curve fitting engines with a functionality ceiling
- New architectures and approaches needed for next generation of machine learning problems to deliver real AI solutions
  - Interconnect technology innovation is key to AI architecture evolution

ARTERIS IP