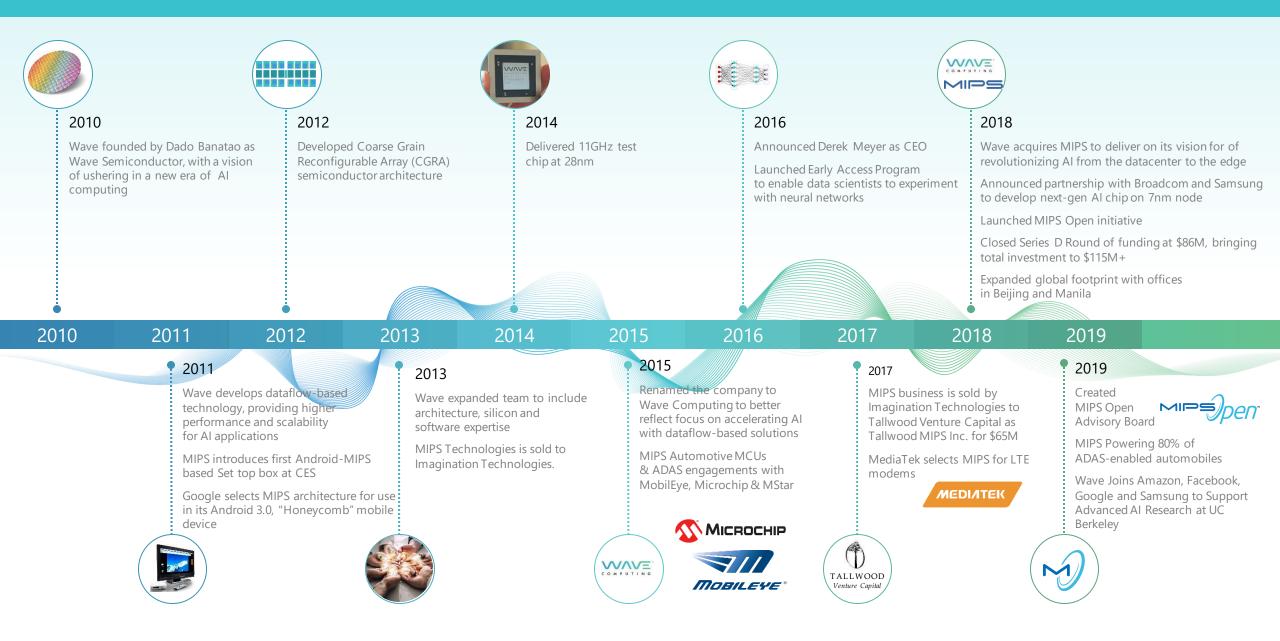
WAVE

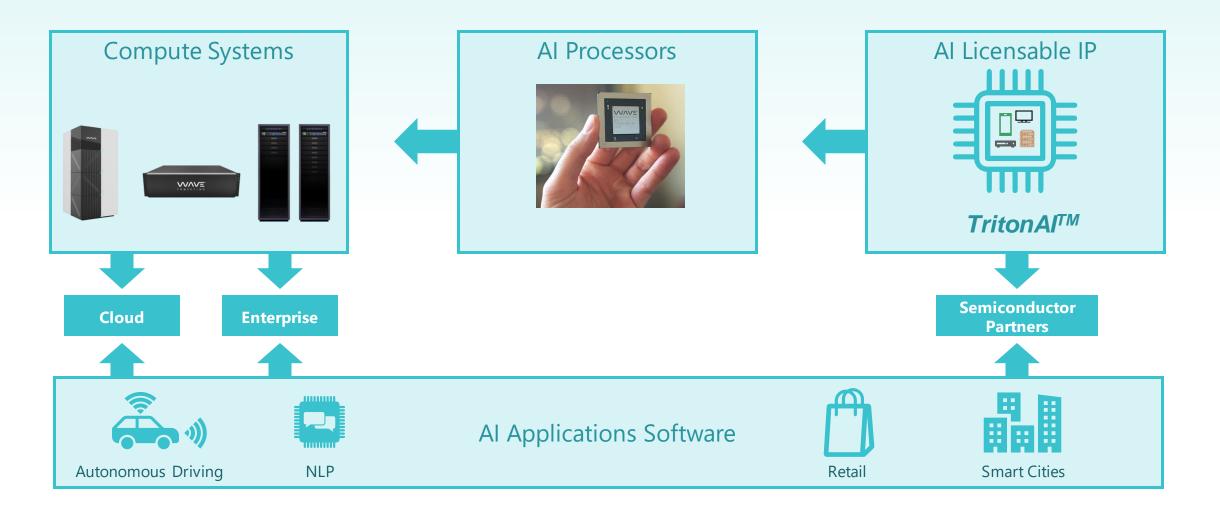
C O M P U T I N G Revolutionizing AI from the Datacenter to the Edge

Adapting the Wave Dataflow Architecture to a Licensable AI IP Product

Presented by **Yuri Panchul**, MIPS Open Technical Lead On SKOLKOVO Robotics & Al Conference. April 15-16, 2019 <u>www.wavecomp.ai</u>

Wave + MIPS: A Powerful History of Innovation







What is Driving AI to the Edge?

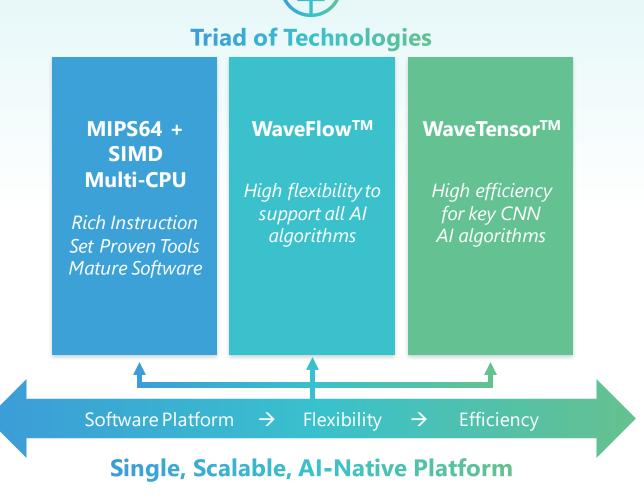




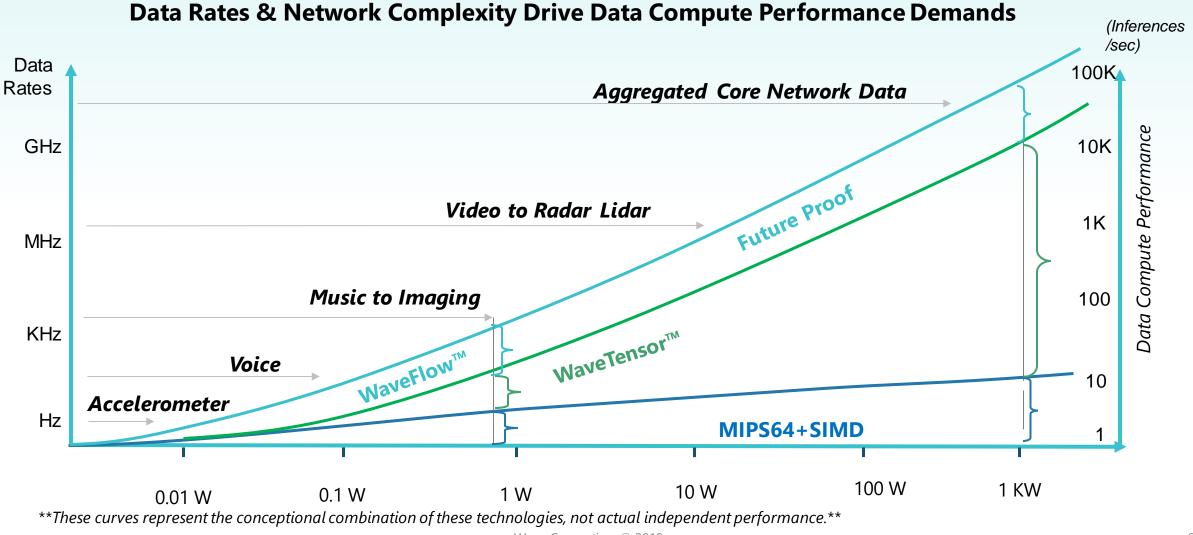
Scalable AI Platform: TritonAI[™] 64

Key Benefits:

- Highly Scalable to address broad AI use cases
- Supports Inference and Training
- High flexibility to support all AI algorithms
- High efficiency for key AI CNN algorithms
- Configurable to support AI use cases
- Mature Software Platform support





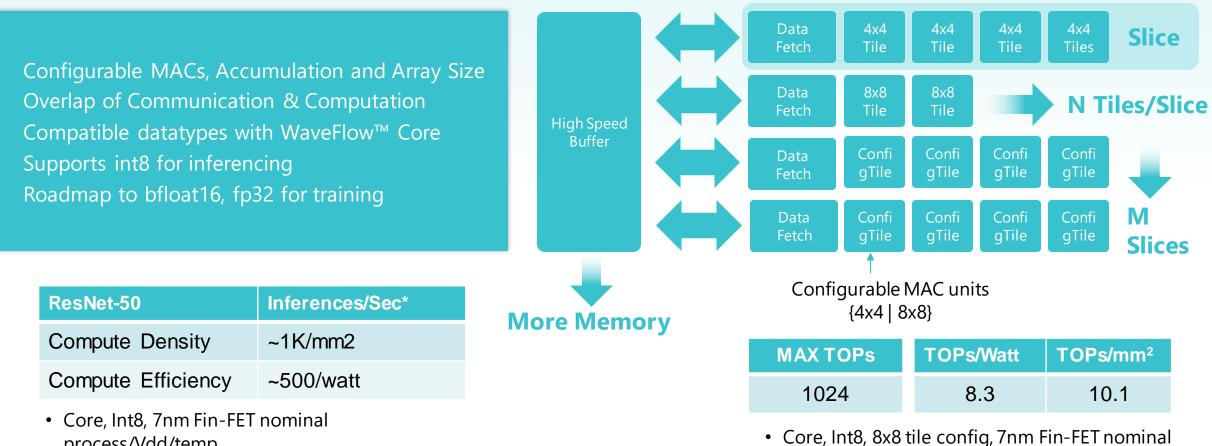


VAVE PUTING

WaveTensor [™] Configurable Architecture

process/Vdd/temp

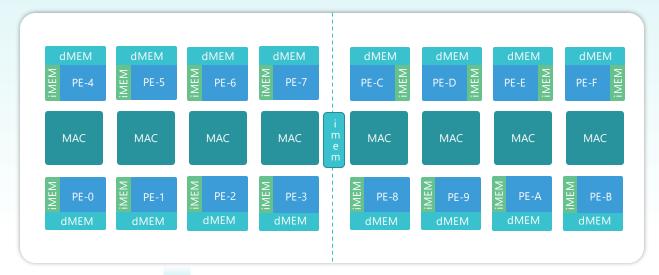
Configurable Architecture for Tensor Processing



- process/Vdd/temp
- Batch=1, std model w/o pruning, performance and power vary with array size/configuration

WaveFlow[™] Reconfigurable Architecture

- Configurable IMEM and DMEM Sizes
- Overlap of communication & Computation
- Compatible datatypes with WaveTensor™
- Integer (Int8, Int16, Int32) for inference
- Roadmap (bfloat16, fp32) for training



Tile = (16 PE's + 8 MACS)

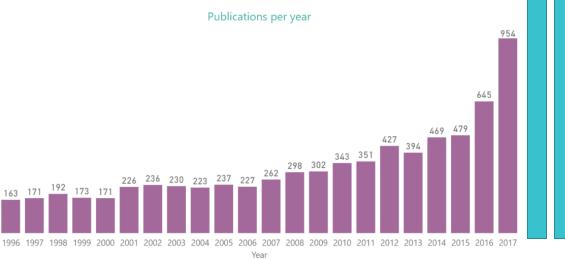
WaveFlow[™] = Wave Dataflow Array of Tiles

- Wide range of scalable solutions (2-1K tiles)
- Future Proof all AI algorithms
- Flexible 2 dimensional tiling implementation
- Reconfigurable for dynamic networks
- Concurrent Network execution
- Supports signal and vision processing

Looks like NIPS 2018 may have sold out in under 15 minutes. For those debating ML hype, getting a ticket to a ML conference is now more challenging than a Taylor Swift conference or a Hamilton showing

Follow

8:22 AM - 4 Sep 2018 from Iceland



What is the likelihood that your DNN accelerator will run all these "yet to be invented" networks?

Wave's TritonAI[™] 64 platform combines a reconfigurable processor with an efficient neural network accelerator.

Offers customers peace of mind and investment protection

Future-proof your Silicon

CNN Layers

- Sparse Matrix-Vector Processing
- Stochastic pooling
- Median pooling (illumination estimation & color correction)

Activation functions

- Leaky rectified linear unit (Leaky ReLU) (used in Yolo3)
- Parametric rectified linear unit (PReLU)
- Randomized leaky rectified linear unit (RReLU)

Custom Operators (e.g.)

- Novel Loss Function
- New Softmax Implementation
- Image resize nearest neighbor

Data Preprocessing

- Scaling
- Aspect Ratio adjustment
- Normalizing

Other Functions

- Compression/Decompression
- Encryption/Decryption
- Sorting

MIPS-64 Processor

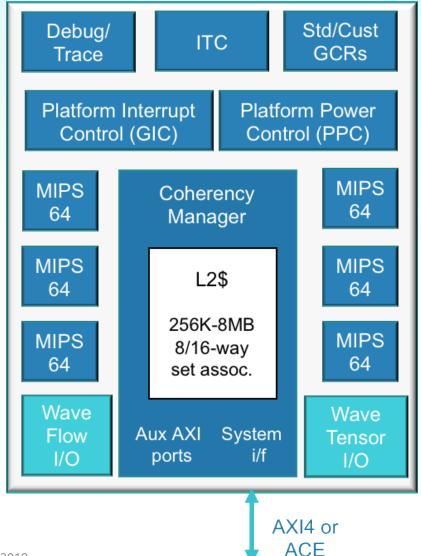
<u>MIPS-64</u>:

• MIPS64r6 ISA

- 128-bit SIMD/FPU for int/SP/DP ops
- Virtualization extensions
- Superscalar 9-stage pipeline w/SMT
- Caches (32KB-64KB), DSPRAM (0-64KB)
- Advanced branch predict and MMU

Multi-Processor Cluster:

- 1-6 cores
- Integrated L2 cache (0-8MB, opt ECC)
- Power mgmt. (F/V gating, per CPU)
- Interrupt control with virtualization
- 256b native AXI4 or ACE interface



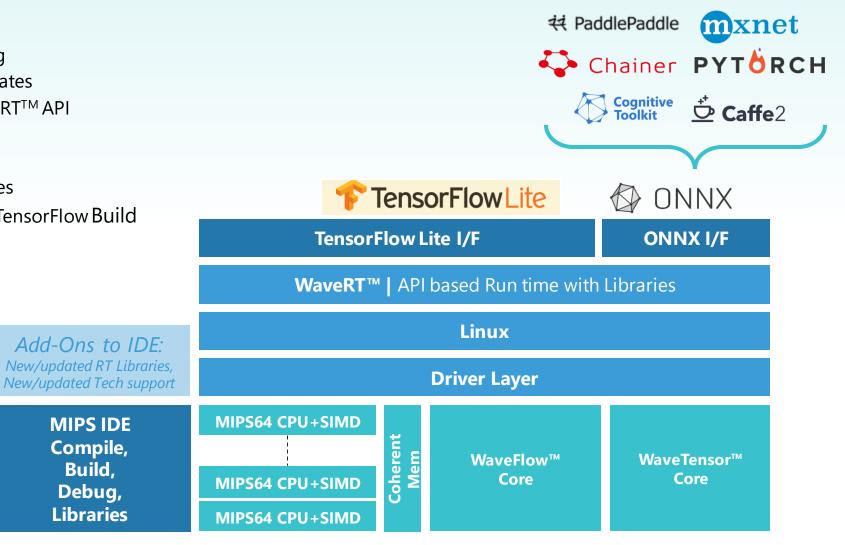
Wave's TritonAI[™] 64 IP Software Platform

Software Platform:

- Mature IDE & Tools
- Driver Layer for Technology Mapping
- Linux Operating system support/updates
- Abstract AI Framework calls via WaveRT[™] API
- **Optimized AI Libraries for:** •
- CPU/SIMD/WaveFlow/WaveTensor
- TensorFlow-Lite Build support/updates
- Extensible to Edge Training with Full TensorFlow Build

Configurable Hardware Platform:

- MIPS64r6 ISA Cluster
 - 1-6 cores
 - 1-4 threads/core
 - L1 I/D (32KB-64KB)
 - Unified L2 (256K to 8 Mbytes)
- WaveFlow Tile Array
 - 4 N Tiles
- WaveTensor Slice Array
 - 1 N Slices



MIPS IDE

Compile,

Build.

Debug,

Libraries

Federated Learning: The Next Frontier in Edge Al

Better ML comes at a cost of collecting data Most training done in the cloud. i.e. Send your data to the cloud.



Diminished Privacy

- Where is your data?
- Who has access to your data?

Incompatible with Banks, Insurance, Military, Health sectors

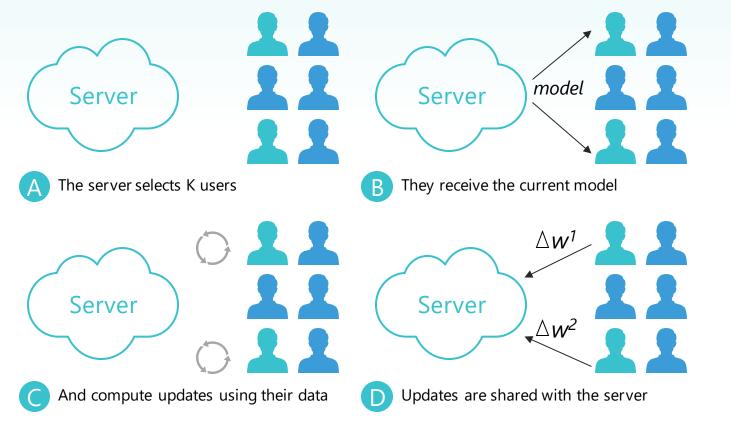
Latency Problems

• Most access technologies are asymmetric

High Communications Costs



Federated learning uses training at the edge to refine the global model

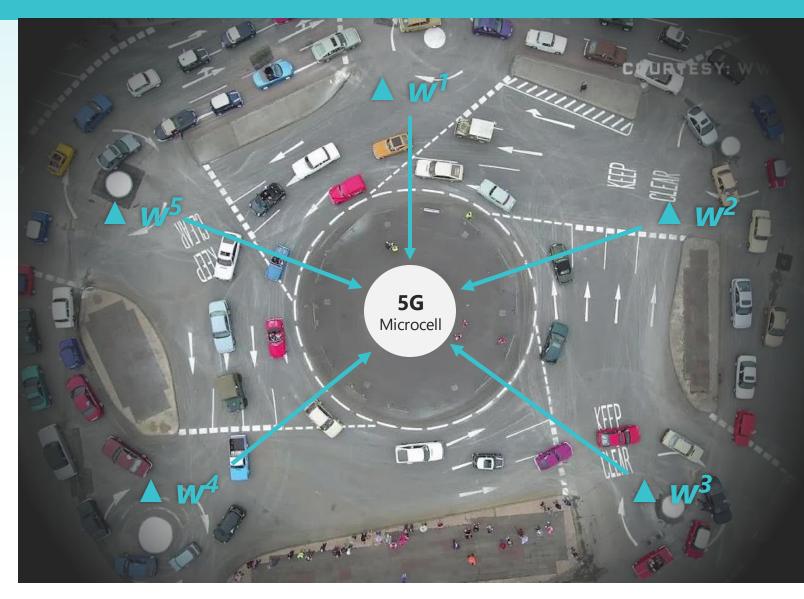


- 1. Server selects a group of users
- 2. Users receive copy of central model
- 3. Users update model based on local data ("Training at the Edge")
- 4. Updates are shared with the server (User data remains private)
- 5. Server aggregates the changes and updates the central model

Federated Learning: Edge AI with Data Privacy

Benefits & Use Cases:

- Transfer learning using local data at edge
- Edge data remains private
- Social networking applications
- Intelligent transportation systems that help increase passenger & pedestrian safety + traffic flow

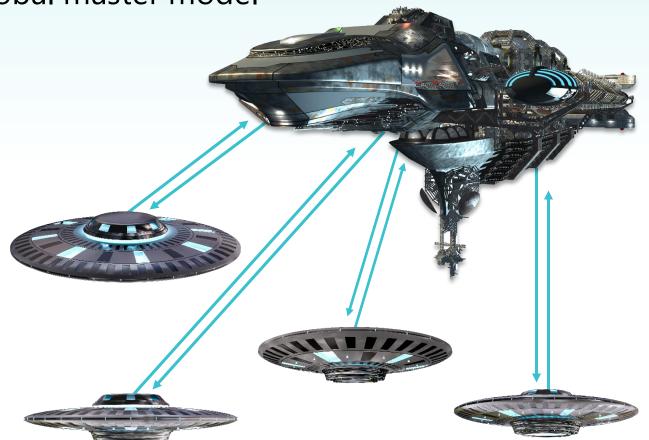




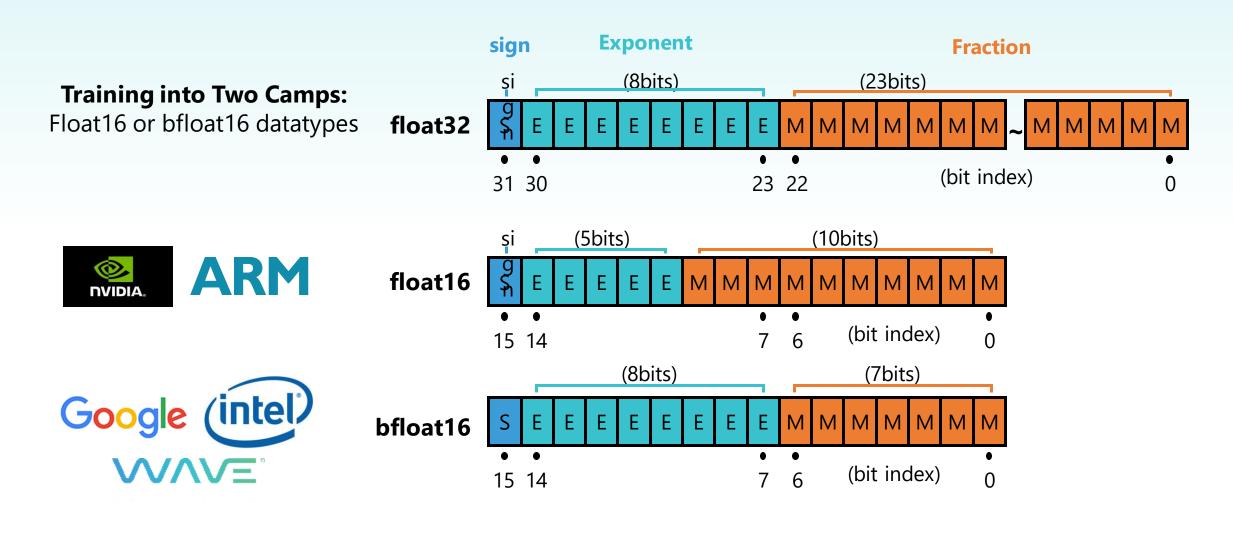
Federated learning uses training at the edge to refine a global master model

Benefits & Use Cases:

- Transfer learning using local data at edge
- Edge data remains private
- Social networking applications
- Intelligent transportation systems that help increase passenger & pedestrian safety + traffic flow

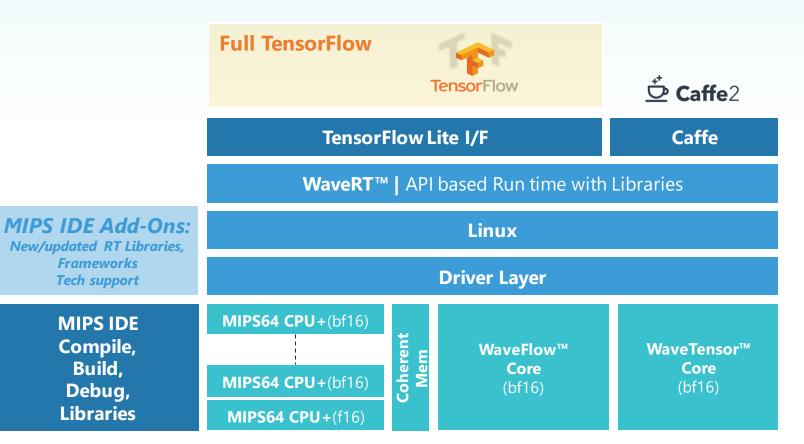






Edge Training Development

- Training Stacks for
 - Federated Learning at the Edge
 - Transfer Learning at the edge
 - Local or personalized models
- Full TensorFlow Build
 - WaveRT API Ext for Training
 - Optimized SIMD FP32 & bfloat16 eigen libraries
 - Deploy training at the edge







Wave's TritonAI[™] Platform Drives Inferencing to the Edge

Wave's TritonAI[™] Platform is a configurable, scalable & programmable offering customers' efficiency, flexibility and AI investment protection

Wave will enable "Training at the edge" with next-gen MIPS AI processor bfloat16 architectures



Thank You

If you have questions or would like more information, visit <u>www.wavecomp.ai</u>



@wavecomputing



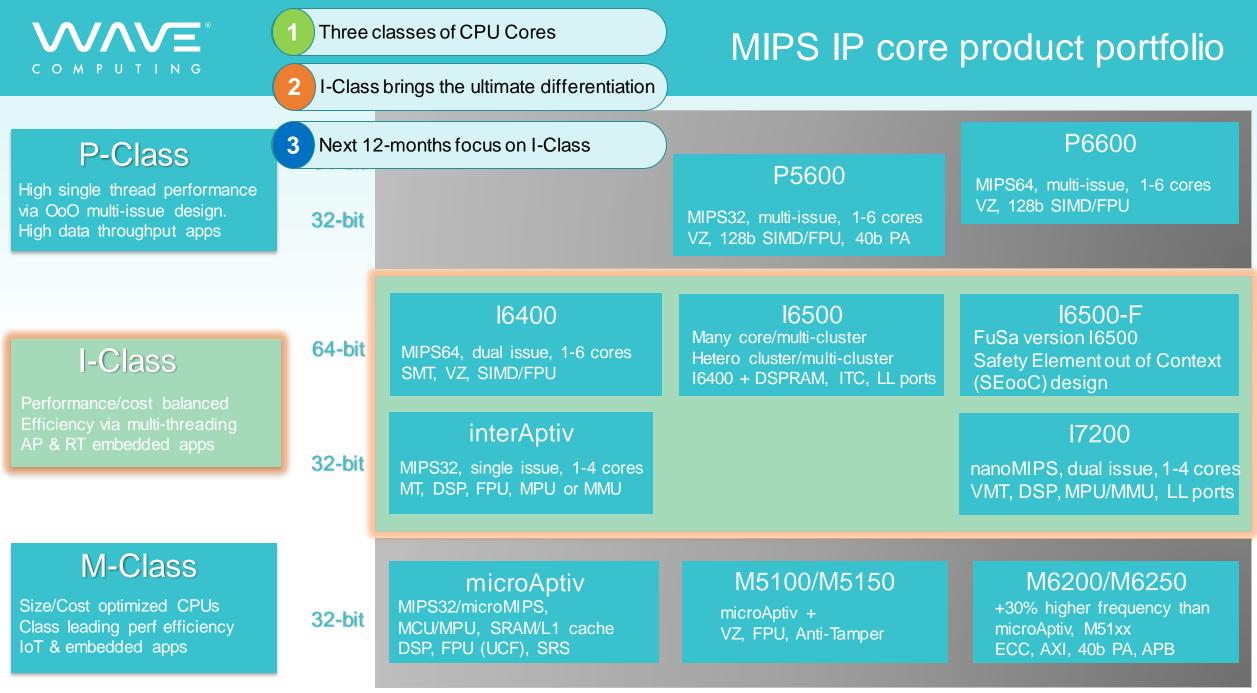
https://www.linkedin.com/company/wave-computing



https://www.facebook.com/WaveComp/

MIPS-Classic Cores

Presented by Yuri Panchul MIPS Open Technical Lead MIPS Open Meetup in Moscow April 15, 2019



MIPS IP Cores - Features Summary

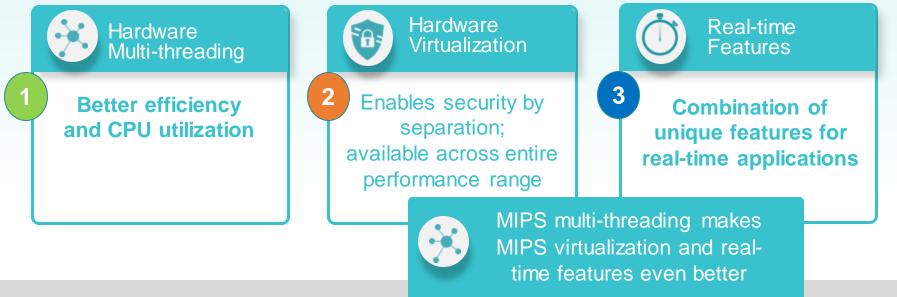
	microAptiv	M51xx	M62xx	interAptiv	I7200	16500/-F	P5600	P6600
MIPS Primary ISA	MIPS32 r5	MIPS32 r5	MIPS32 r6	MIPS32 r5	nanoMIPS32	MIPS64 r6	MIPS32 r5	MIPS64 r6
Virtual/Phys Addr Bits	32/32	32/32	32/32	32/32	32/32	48/48	32/40	48/40
FPU	\checkmark (UC version only)	\checkmark	-	MT	-	MT w/SIMD	Hi Perf w/SIMD	Hi Perf w/SIMD
DSP/SIMD extensions	DSPASE r2	DSPASE r2	DSPASE r2	DSPASE r2	DSPASE r2	MSA 128-bit	MSA 128-bit	MSA 128-bit
Virtualization	-	\checkmark	-	-	-	\checkmark	\checkmark	\checkmark
Small code size ISA	microMIPS32	microMIPS32	microMIPS32	MIPS16e2 ASE	nanoMIPS32	-	-	-
Multi-threading	-	-	-	2 VPE, 9 TC	3 VPE, 9 TC	4 VPE	-	-
SuperScalar	-	-	-	-	Dual-issue in order	Dual-issue in order	Multi-issue OoO	Multi-issue OoO
Pipeline stages	5	5	6	9	9	9	16	16
Relative Frequency*	0.6x	0.6x	0.75x	1x	0.95x	0.90x	1.10x	1.10x
SPRAMs (I/D/U)	✓ / ✓ / -	√ / √ / -	✓ / ✓ / -	√ / √ / -	$\checkmark / \checkmark / \checkmark$	- / ✓ / -	- / - / -	- / - / -
L1 caches	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark
L2 cache	-	-		\checkmark	\checkmark	\checkmark	\checkmark	\checkmark
Coherent Multi-Core	-	-	-	Up to 4 cores	Up to 4 cores	Up to 6 cores, Up to 64 clusters	Up to 6 cores	Up to 6 cores
Native System Bus I/F	AHB-Lite	AHB-Lite	AXI	OCP 2 or AXI	AXI	AXI or ACE	AXI	AXI

* Relative Frequencies are approximate, are provided for rough guidance only, and will vary to some extent in different process nodes



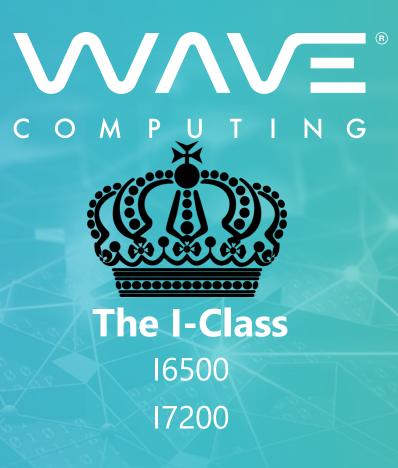
MIPS technology differentiation

MIPS architecture and IP cores offer powerful, unique capabilities



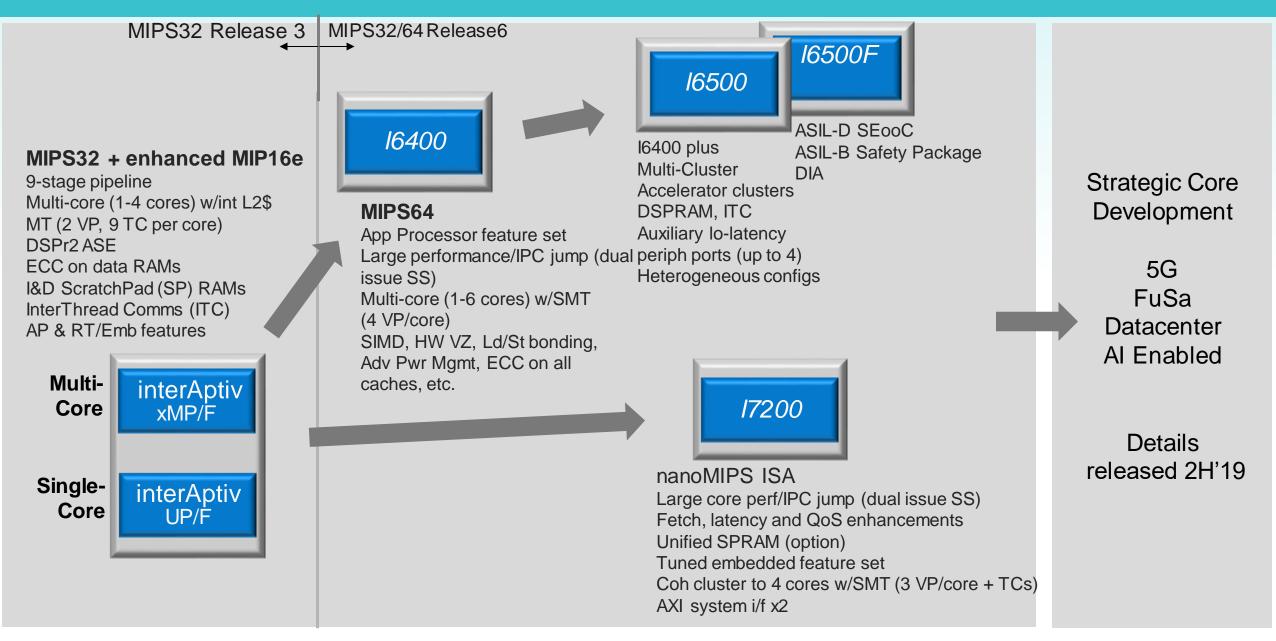
MIPS IP cores

- 1. Offer leading Power Performance Area (PPA) across the range
- 2. Provide ultimate scalability: multi-thread, multi-core, multi-cluster
- 3. Address Functional Safety to ISO 26262 for automotive and IEC 61508 for industrial



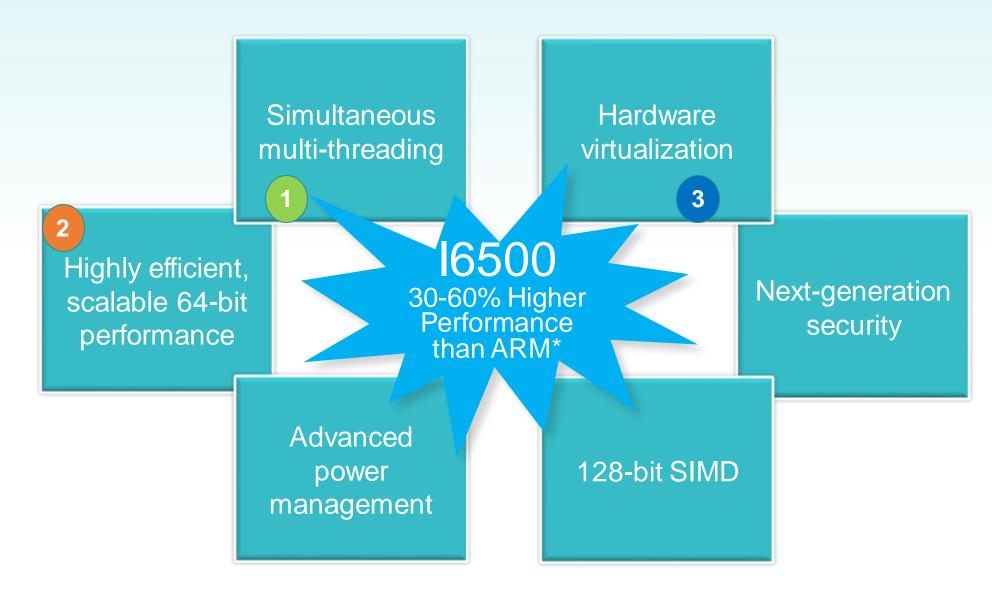
MIPS® I-Class Processor Core Roadmap

MIPS32 and MIPS64 evolution



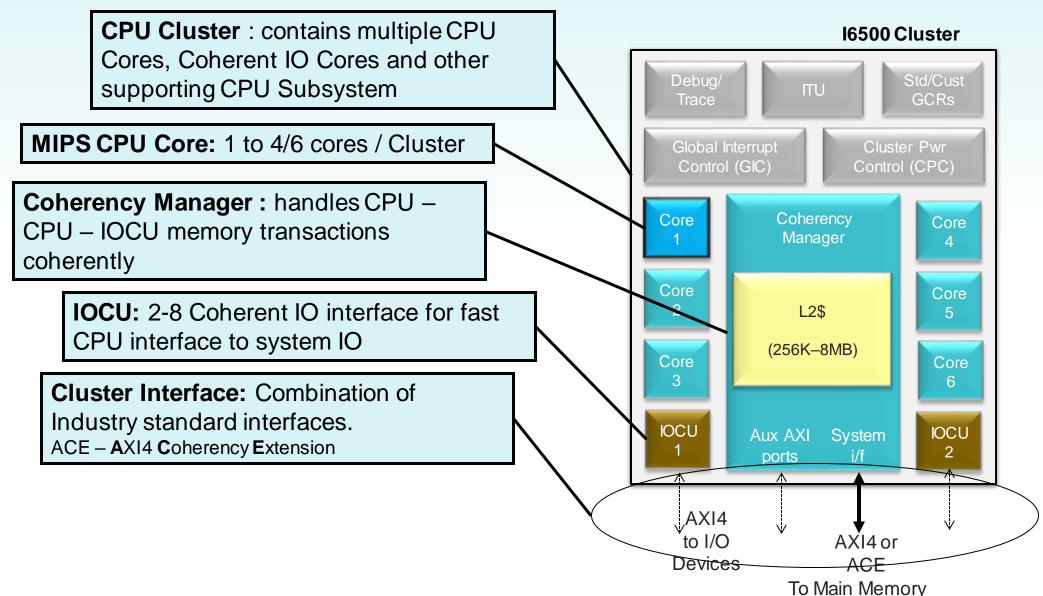


Setting a new standards in mainstream 64-bit processing



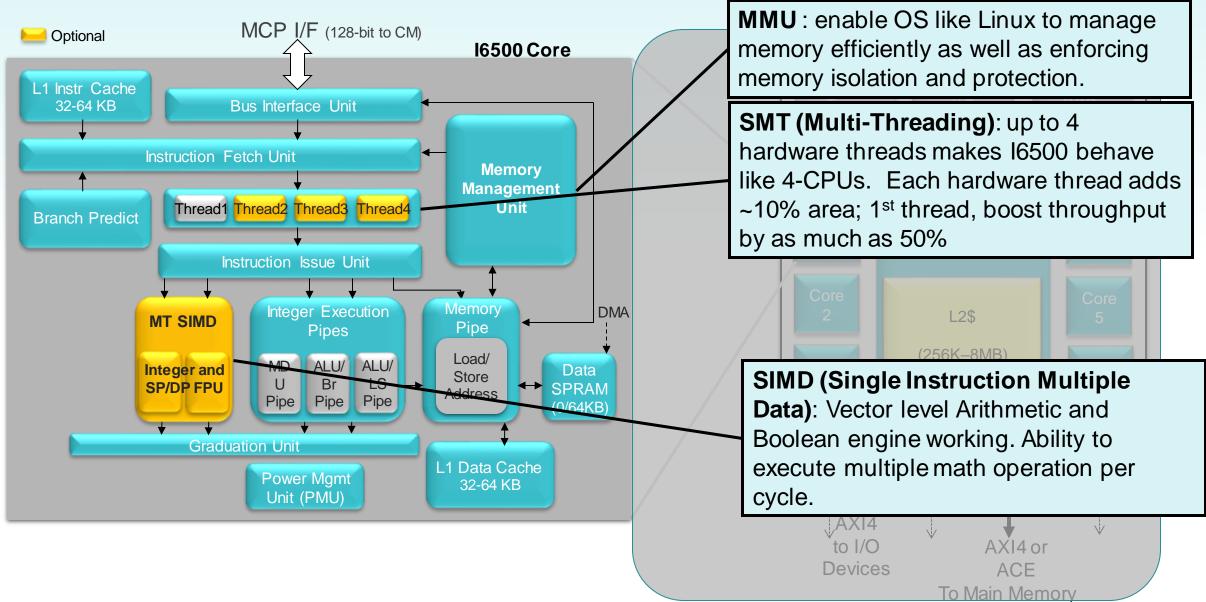
Key Components of a Multi-Core Cluster?



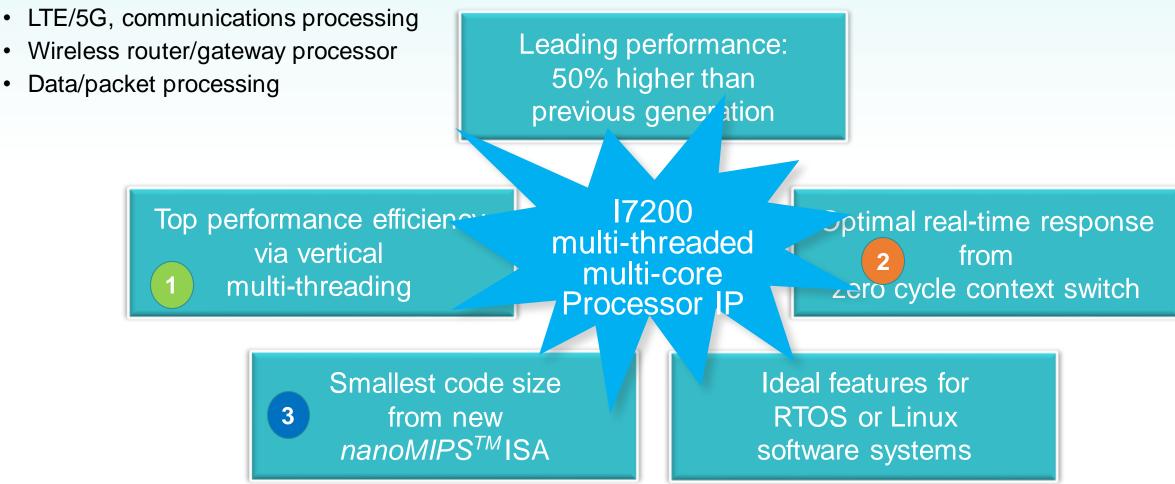


How to build a Multi-Core Cluster?

16500



- Multi-threaded multi-core 32-bit processor IP
- Designed for high performance embedded systems with real-time requirements

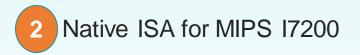


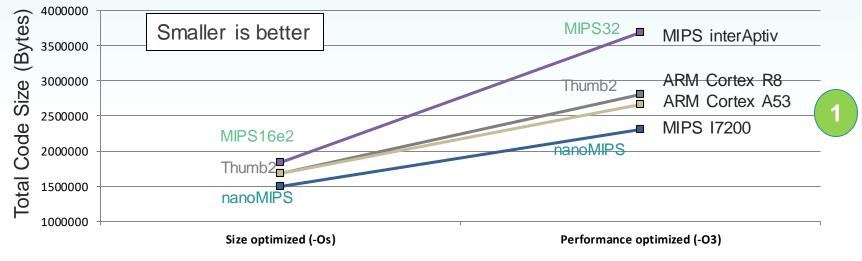


3

nanoMIPS ISA – advancing small code size

Achieves outstanding code density targeting small footprint applications or constraint devices





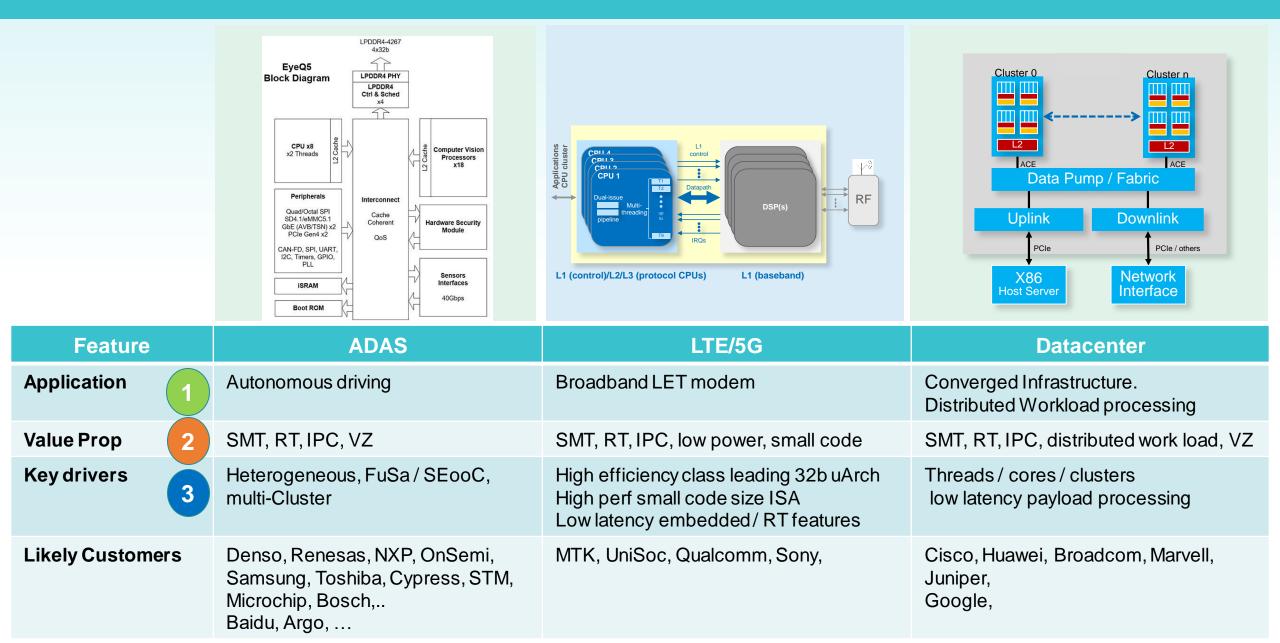
gcc compiler, with indicated optimization target



Success Stories

Market specific use-cases

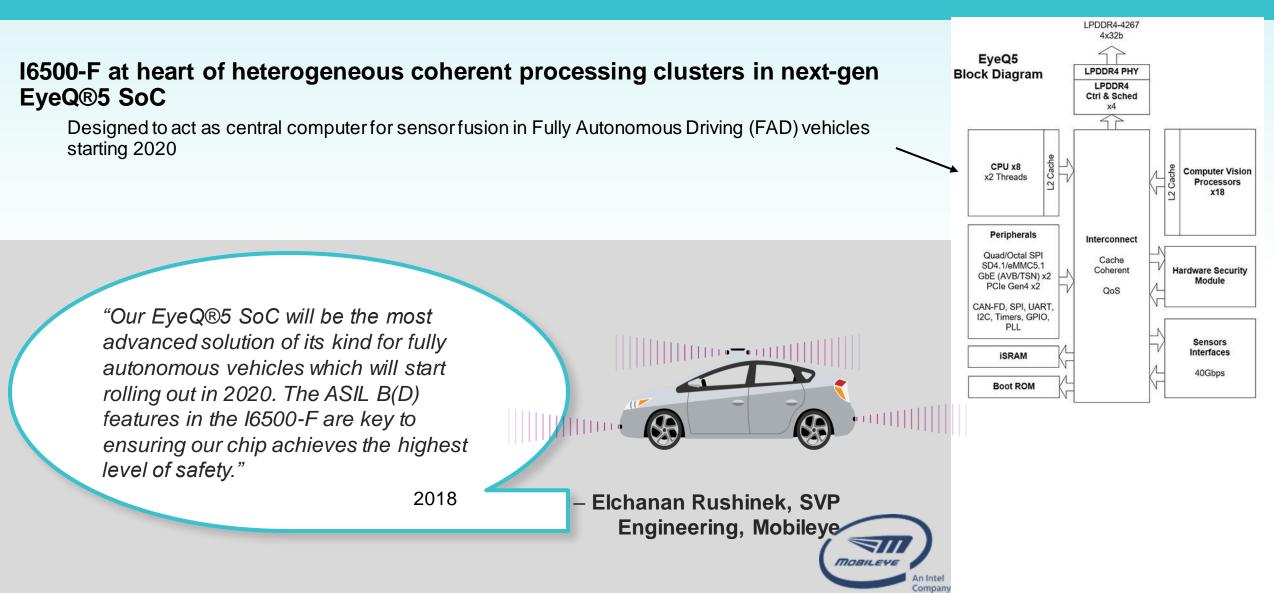
Classical Application Segments





Application Segment 1 ADAS

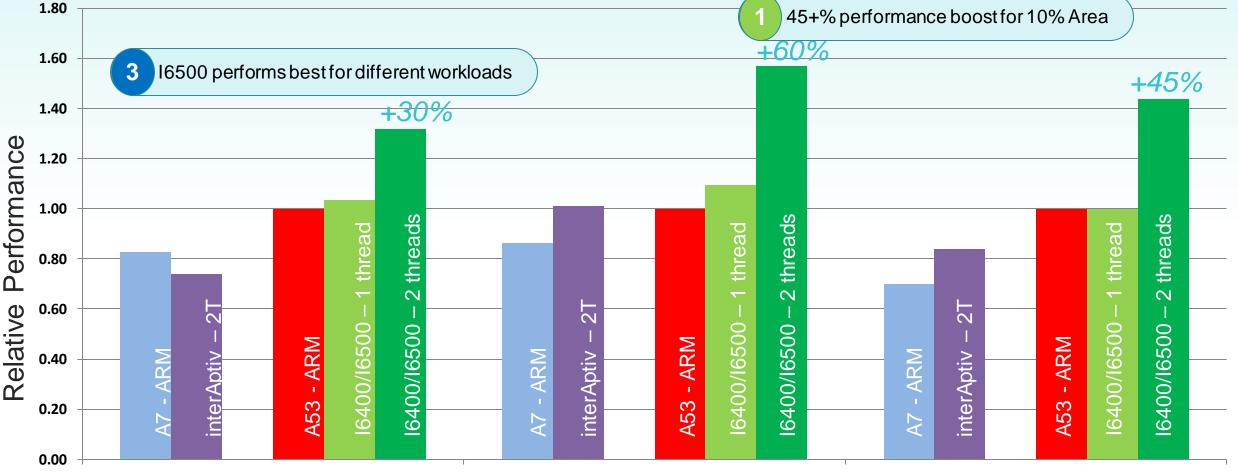
I6500-F lead customer: Mobileye ADAS Platform





Per core, normalized to A53 Values (@ same frequency)

Simultaneous Multi-Threading



DMIPS - 32b

CoreMark

SPECint2000 (rate)

Based on Cortex A53 data reported by ARM on website and/or in presentation materials, plus benchmarked results on Linaro (HiSilicon Octa A53 Kirin620) with Linux kernel: 3.18.0-linaro-hikey SMP preempt, RFS Debian squeeze, with GCC-based 5.0.0 toolchain. A7 scores are ARM claims. Measured results are lower.

16400 results are based on production released RTL, FPGA platform benchmarking and in case of SPEC, 1 enhancement for next release performance models • testing



Application Segment 2 LTE/5G

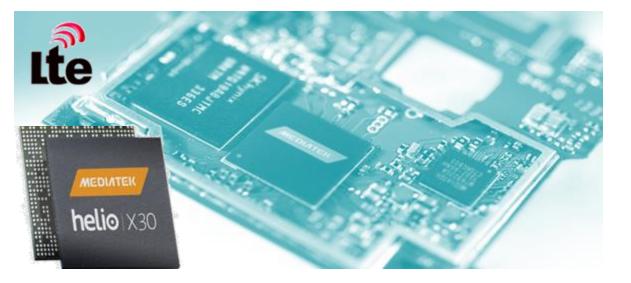


Success Story, Multi-threading & Mediatek

17200

RF

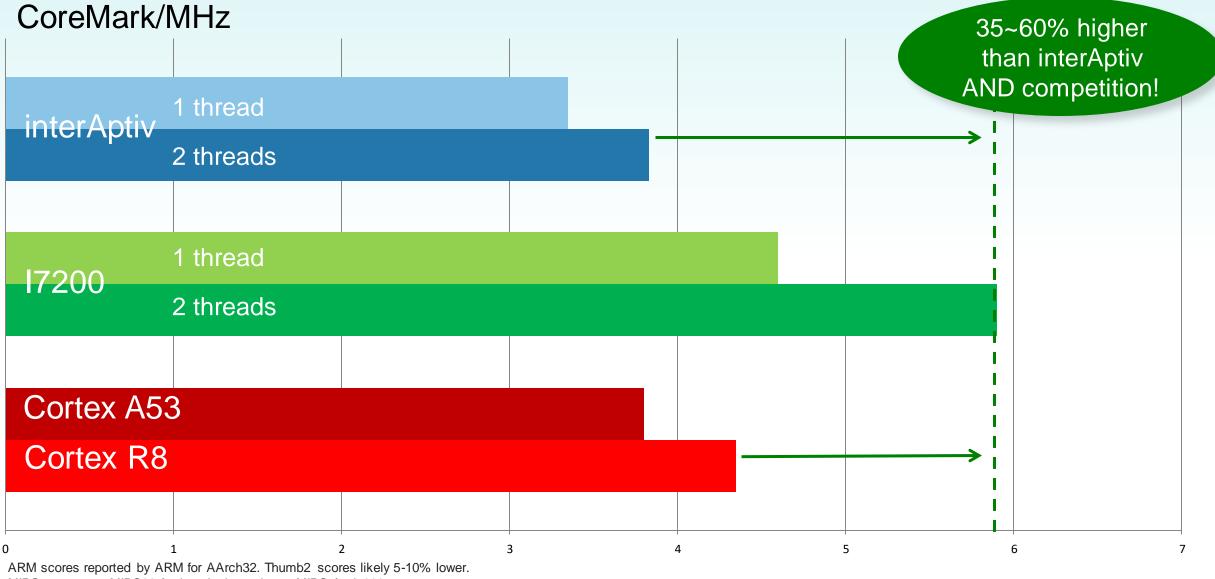
	Mediatek 5G LTE modem	About MIPS in the application
LTE	 Mediatek designed MIPS into modem across multiple products First version, Helio X30, in production 	
MIPS Advantage	 Hardware Multi-Threading (MT) Fast inter-thread communications Higher performance/high processing efficiency Scalability 1-4 cores, 2 threads per core 	Stories CPU A CPU 2 CPU 2 CPU 1 CPU 1 CPU 1 CPU 1 CPU 2 CPU 1 Dual-issue threading
	Deterministic ; Real-time interrupts	L1 (control)/L2/L3 (protocol CPUs) L1 (baseband)



"MIPS CPUs, with their powerful multi-threading capability, offer a combination of efficiency and high throughput for LTE modems that contributes significantly to system performance." *said Dr. Kevin Jou, SVP and CTO, MediaTek*. 2018



I7200 Performance Advantage



MIPS scores are MIPS32 for interAptiv, and nanoMIPS for I7200



Application Segment 3 Data-center



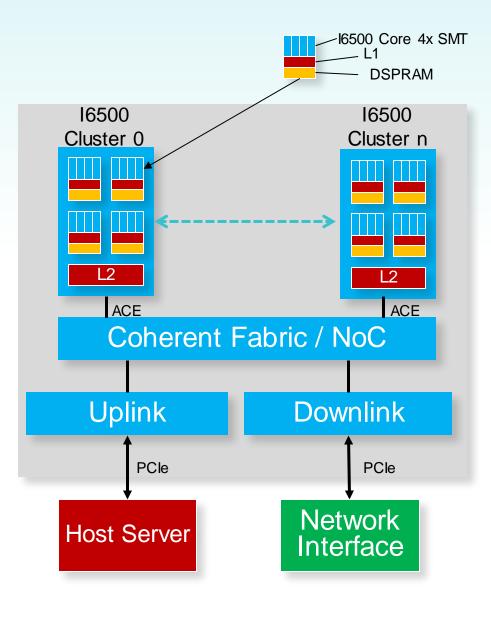
Scala

3

- Scalable Multi-threading to Multi-core to Multi-cluster
- Configurable Heterogeneous inside & outside
- Optimized for High-throughput data processing applications
- Real-time, secure, deterministic and low latency

"The MIPS Simultaneous Multi-Threading architecture is an important technique to ensure that such workloads run efficiently as measured by the CPU's instructions per clock, or IPC. We have seen that this efficiency translates directly into a smaller area as well as lower power for silicon implementations based on MIPS." 2018

Pradeep Sindhu, CEO of Fungible



VAVE

16400 & 6500 – Why Multi Threading?

A powerful differentiator among CPU IP cores



3

Why MT?

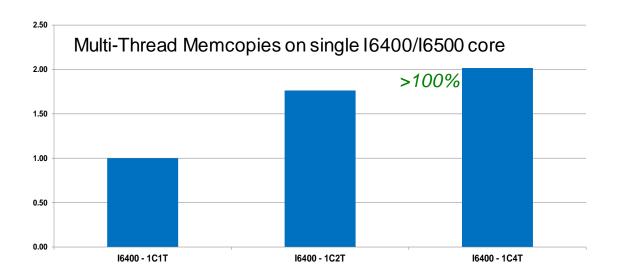
- A path to higher performance, and higher efficiency
- 30%-60% higher performance for 10% per Thread increase

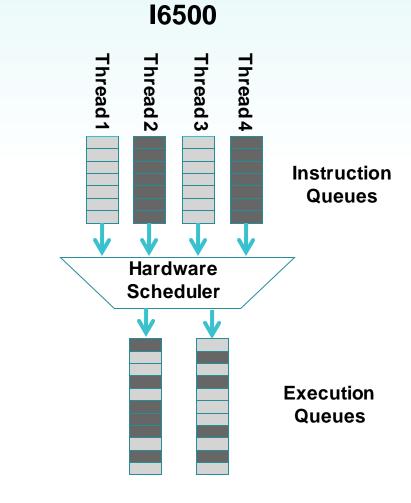
Easy to use – programming model is same as multi-core

A thread looks like a core to standard SMP OS

Simultaneous / concurrent execution

Zero Cycle overhead context switching





MIPS IP Cores – Mapping to ARM

	microAptiv	M51xx	M62xx	interAptiv	I7200	16500/-F	P5600	P6600
MIPS Primary ISA	MIPS32 r5	MIPS32 r5	MIPS32 r6	MIPS32 r5	nanoMIPS32	MIPS64 r6	MIPS32 r5	MIPS64 r6
Virtual/Phys Addr Bits	32/32	32/32	32/32	32/32	32/32	48/48	32/40	48/40
FPU	\checkmark (UC version only)	\checkmark	-	MT	-	MT w/SIMD	Hi Perf w/SIMD	Hi Perf w/SIMD
DSP/SIMD extensions	DSPASE r2	DSPASE r2	DSPASE r2	DSPASE r2	DSPASE r2	MSA 128-bit	MSA 128-bit	MSA 128-bit
Virtualization	-	\checkmark	-	-	-	\checkmark	\checkmark	\checkmark
Small code size ISA	microMIPS32	microMIPS32	microMIPS32	MIPS16e2 ASE	nanoMIPS32	-	-	-
Multi-threading	-	-	-	2 VPE, 9 TC	3 VPE, 9 TC	4 VPE	-	-
SuperScalar	-	-	-	-	Dual-issue in order	Dual-issue in order	Multi-issue OoO	Multi-issue OoO
Pipeline stages	5	5	6	9	9	9	16	16
Relative Frequency*	0.6x	0.6x	0.75x	1x	0.95x	0.90x	1.10x	1.10x
SPRAMs (I/D/U)	✓ / ✓ / -	√ / √ / -	✓ / ✓ / -	✓ / ✓ / -	$\checkmark \checkmark \checkmark$	- / ✓ / -	- / - / -	- / - / -
L1 caches	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark
L2 cache	-	-		\checkmark	\checkmark	\checkmark	\checkmark	\checkmark
Coherent Multi-Core	-	-	-	Up to 4 cores	Up to 4 cores	Up to 6 cores, Up to 64 clusters	Up to 6 cores	Up to 6 cores
Native System Bus I/F	AHB-Lite	AHB-Lite	AXI	OCP 2 or AXI	AXI	AXI or ACE	AXI	AXI
Sample ARM Mapping	M3 I	M4	M23 M33	uting Confidential © 20	R52 R7 R8	A53	A57	A72



MIPS Open™ The New Standard in Open Use ISAs

Presented by Yuri Panchul, MIPS Open Technical Lead MIPS Open Meetup in Moscow, April 15, 2019

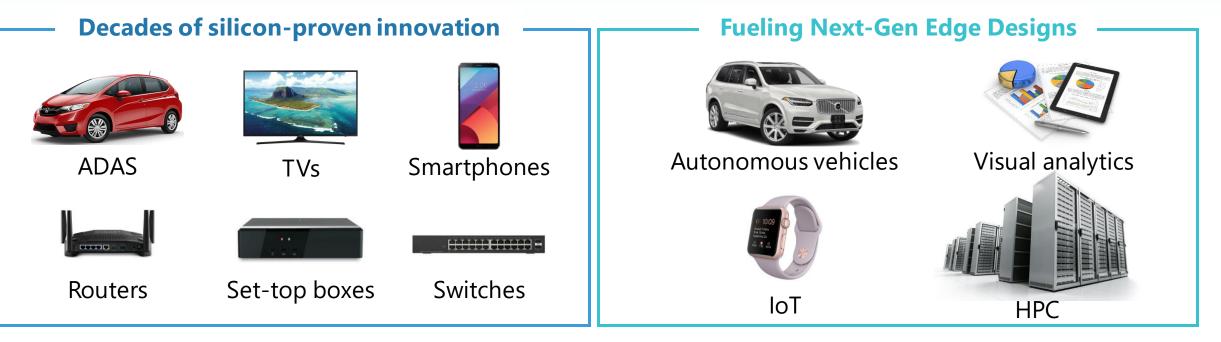


Wave Computing © 2019



mips MIPS Computer Systems Inc. founded 1981		IF	PS becomes Plicensing company			Smar	<mark>рилтек</mark> tphone hodem		Wave Establishes MIPS Open Advisory Board
1980		1990		2000	2	010	2017	2018	2019
1981 MIPS architecture invented by team from StanfordImage: stanford	First MIPS product ships: R2000 uP	DTV/STB product ships: NEC SONY	Cable Modem product ships: TOSHIBA	Wi-Fi - Residential Gateway	<section-header></section-header>	Automo MCUs & ADAS MICI MICI	S ROCHIP	Wave Acquires MIPS Wave Launches MIPS Open	Joins AWS, Facebook, Google & Samsung on UC Berkeley's BAIR initiative

- Accelerates innovation for system-on-chip designs
- Expands adoption of MIPS RISC architecture & grows supporting ecosystem
- Dedicated Advisory Board prevents architectural fragmentation
- Delivers peace of mind & investment protection for existing MIPS customers

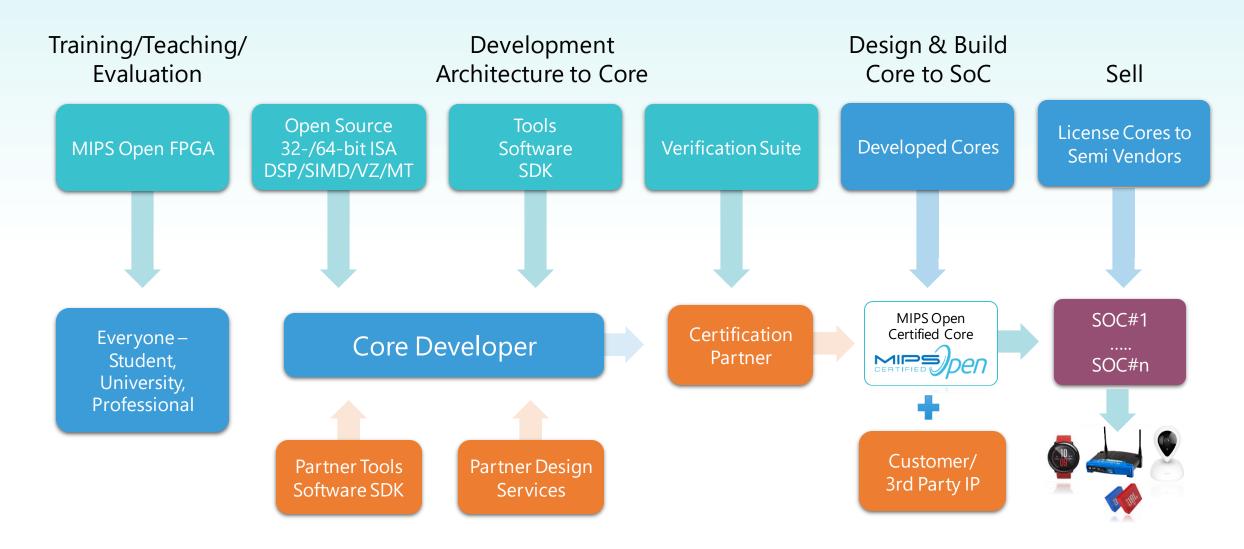




Unified Approach	Proven	Established	Matured
	Architecture	Ecosystem	Technology
 Compatibility No Fragmentation Same tools, software Flexibility – UDI support 	 MIPS architecture- based designs are shipping in billions of devices Shipping Products – wired modems, wireless modems – LTE/5G/Wi-Fi, IoT, automotive – ADAS, data centers, AI and others. 	 Well-established Tools, software, applications, boards Supported and used by customers/ partners 	 30+ years of legacy of MIPS architecture Enabler to build the next trillion devices/ applications

MIPS Open is the World's First Commercial-ready, Silicon Proven, Feature-Rich Architecture Available for Open Use





C O M P U T I N G

MIPSOpen.com Now Live



DOWNLOADS MIPS OPENTM COMPONENTS V RESOURCES V ABOUT V



Wave's First MIPS Open Program Components Now Live

Immediate Access to the Proven, Industry-Standard and Patent-Protected MIPS RISC Architecture





Stave Computing © 2019 – Confidential

MIPS Open[™] Components

MIPS Open ISA	MIPS Open Tools	MIPS Open FPGA	MIPS Open Cores
 Latest R6 version of 32- bit/64-bit ISA Extensions such as Virtualization, Multi- threading, SIMD, DSP and microMIPS architectures 	 IDE for Embedded RTOS and Linux Embedded Edition Enables MIPS software developers to build, debug, and deploy applications on MIPS-based hardware and software platforms 	 Getting Started Guide - provides MIPS Open FPGA system as a set of Verilog files Labs - includes 25 hands-on labs that guide users in exploring computer architecture & system-level design SOC - shows how to build a system-on-chip design based on MIPS Open FPGA that loads the open source Linux 	 Low power, low footprint microAptiv cores Microprocessor(MPU) Microcontroller (MCU) Targeted for embedded applications

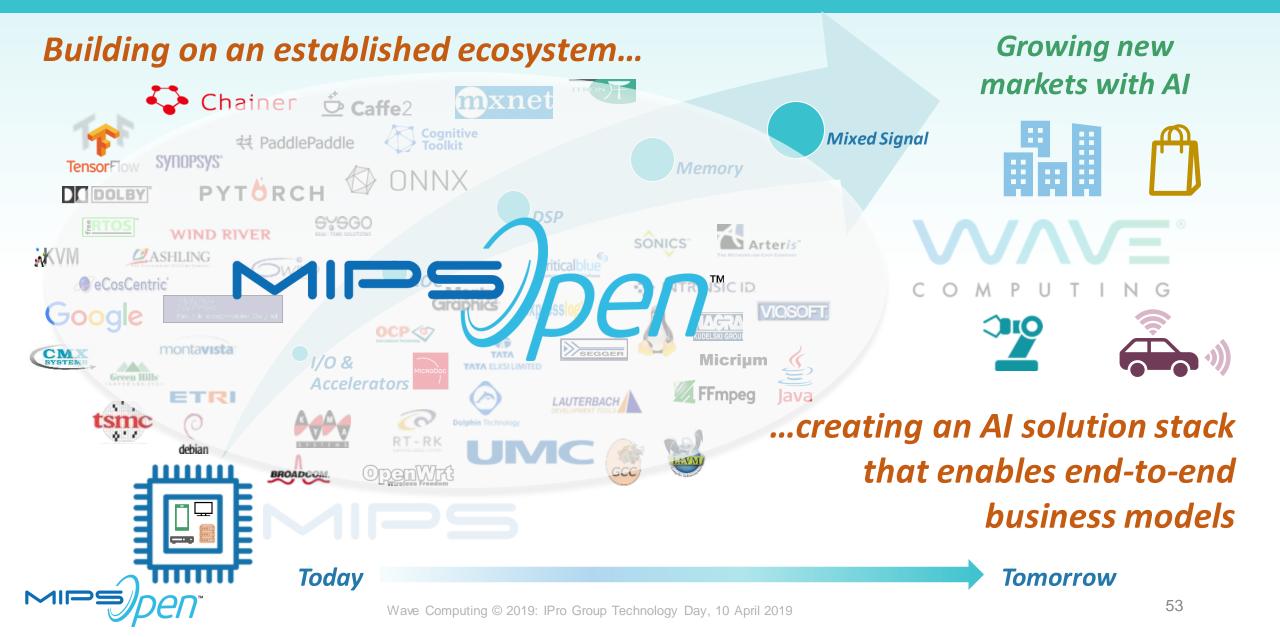
MIPS Open Components includes the latest MIPS R6 ISA, low power, low footprint microAptiv cores, microAptiv soft core targeted for FPGA and tools for software development.



Open Use License	Certification	Patent License	Comprehensive Package
 MIPS Open use license includes: Latest MIPS R6 architecture documents microAptiv cores Tools microAptiv softcore targeted for FPGA 	 Access to certification services to core developers Verification and Certification of cores developed using MIPS R6 ISA. 	 Right and license under R6 architecture patents to design, build and sell cores Use of the "MIPS Certified" trademark logo for certified cores 	 Complete package Instruction set, cores, tools for the community to accelerate innovation at the edge

MIPS Open provides a comprehensive open use package and tools to enable fast time to market for certified cores.

Evolving the MIPS Ecosystem for AI at the Edge



Individual Membership Entry level

FREE

- Allows participation in all working groups and access to all working group materials.
- Represent individual, academic and non profits interests.

Silver Membership Entry level of corporate membership.

\$ 10,000

per year

- Allows participation in all working groups and access to all working group materials.
- Right to vote on working group initiatives.
- Can be appointed to lead a working group. If so appointed, automatically become a member of the technical steering committee.
- Can be nominated to run for election to the board of directors as a Silver class representative.
- Vote as a Silver class for representation on the board of directors.

Gold Membership Middle tier of membership.

\$ 50,000 per year

- Allows participation in all working groups and access to all working group materials.
- Right to vote on working group initiatives.
- Can be appointed to lead a working group. If so appointed, automatically become a member of the technical steering committee.
- Can be nominated to run for election to the board of directors as a Gold class representative.
- Vote as a Gold class for representation on the board of directors. Allows participation in all working groups and access to all working group materials.

Platinum Membership

Decision-making level of membership

\$ 100,000 per year

- Leadership and decision-making level of membership.
- Automatically appointed to the board of directors and the technical steering committee.
- Set direction, approve budgets and projects, appoint working group leaders, create new working groups, and supervise foundation staff.
- Have access to all foundation meetings and materials.
- Allows participation in all working groups and access to all working group materials.
- Right to vote on working group initiatives.Leadership and decision-making level of membership.





Thank You



Wave Computing © 2019