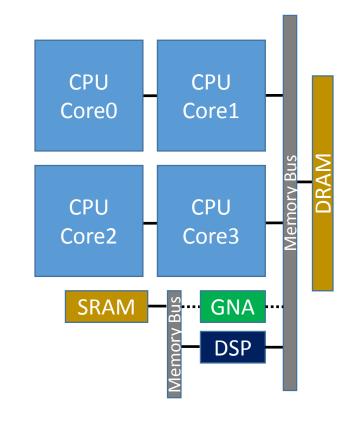
Michael Deisher michael.deisher@intel.com Intel Corporation Hillsboro, Oregon USA





What is GNA?

Low power neural co-processor for <u>continuous inference at the "edge"</u> Designed for Intel[®] Quark[™], Intel Atom[®], and Intel[®] Core[™] based devices Runs while application processor is in low power sleep state Interfaces to system or private memory avoiding CPU cache pollution



How does GNA work?

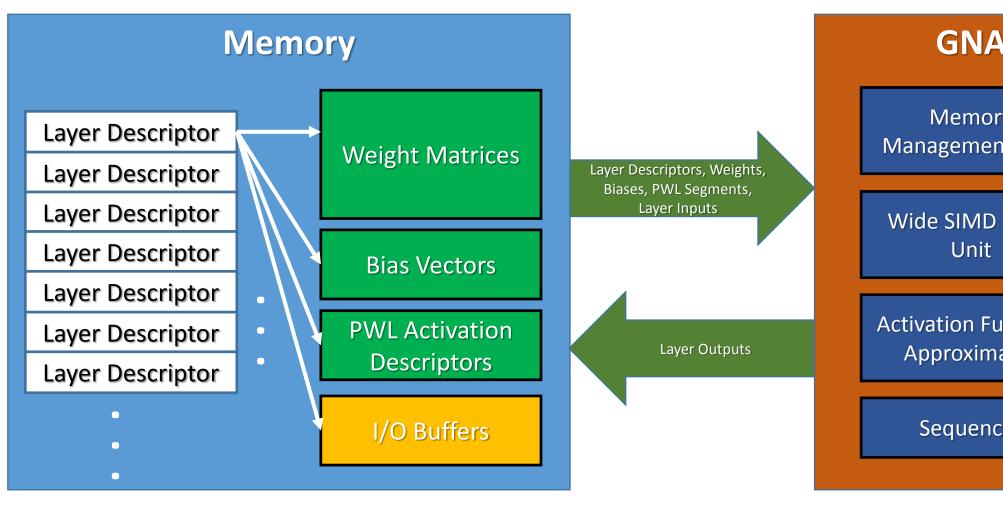
Neural network topology stored in memory as list of layer descriptors Layer types: affine, diagonal affine, Gaussian mixture model, recurrent, convolutional1D, transpose, copy

Activation function: piecewise linear (PWL) approximation

	Input	Output	Batch	Weight	PWL	Partial
Layer Type	Orientation	Orientation	Size	Width	Activation	Output
Affine	column vector	column vector	1-8	1B, 2B	optional	yes
Diagonal Affine	column vector	column vector	1-8	1B, 2B	optional	no
Convolutional	row vector	row vector	1	2B	optional	no
Gaussian Mix	column vector	column vector	1-8	1B	no	yes
Recurrent	row vector	row vector	1-8	1B, 2B	required	no
Сору	row vector	row vector	1-8	N/A	N/A	no
Interleave	row vector	column vector	1-8	N/A	N/A	no
Deinterleave	column vector	row vector	1-8	N/A	N/A	no

All-integer math (inputs, outputs, weights, biases) Batching of layer inputs for better throughput Optional on-the-fly pruning (affine layer) Complex graphs (e.g., LSTM, GRU, TDNN) constructed from Stream processing model

- App processor configures memory, starts GNA, and sleeps or
- GNA signals when forward propagation is complete



IMPLEMENTATION OF EFFICIENT, LOW POWER DEEP NEURAL NETWORKS ON NEXT-GENERATION INTEL CLIENT PLATFORMS



How is GNA used?

Start with floating point neural network trained in framework of choice Import using Intel[®] Deep Learning SDK Deployment Tool or Kaldi example Link with Intel[®] Deep Learning SDK Inference Engine or GNA native library

Neural Network Training Frameworks

Kaldi, Tensorflow, CNTK, Theano, Caffe

Intel DL SDK, GNA Example Code

GNA native library options

- Firmware API (for Intel[®] Quark[™] running real-time operating system)
- Middleware API (for Intel Atom[®] and Intel[®] Core[™] running Linux, Windows)

API Function	Description
GNADeviceOpen	acquire handle to GNA device
GNADeviceClose	release handle to GNA device
GNAAlloc	allocate memory (and pin so it
GNAFree	free GNA memory (after unpin
GNAPropagateForward	propagate inputs through all l
GNAWait	wait until propagation request

Complexity is hidden in layer descriptor list construction

- Handled transparently by Intel[®] Deep Learning SDK model optimizer
- Full control via GNA native library API

mach (mpacs, outputs, weights, blases)	Intel [®] Deep Learning SDK Dep Enables full utilization of IA Inference while abstract		
layer inputs for better throughput			
n-the-fly pruning (affine layer)	Optimize:		
aphs (e.g., LSTM, GRU, TDNN) constructed from basic layer types cessing model essor configures memory, starts GNA, and sleeps or does useful work als when forward propagation is complete	 Imports trained models from all popular DL framework regardless of training HW Model Canonicalization, Compression and Quantization Deploy: 		
Memory Layer Descriptor Layer Descriptor <td< td=""><td> One API across all Intel HW and systems Friendly Inference solution: (low footprint, easy API, control meeting Functional Safety) Optimizes Inference execution per target hardware under-the-hood Ease of use + Embedded friendly + E https://software.intel.com/en-us/en</td></td<>	 One API across all Intel HW and systems Friendly Inference solution: (low footprint, easy API, control meeting Functional Safety) Optimizes Inference execution per target hardware under-the-hood Ease of use + Embedded friendly + E https://software.intel.com/en-us/en		
	orkloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYS my of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluation		

Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other product.

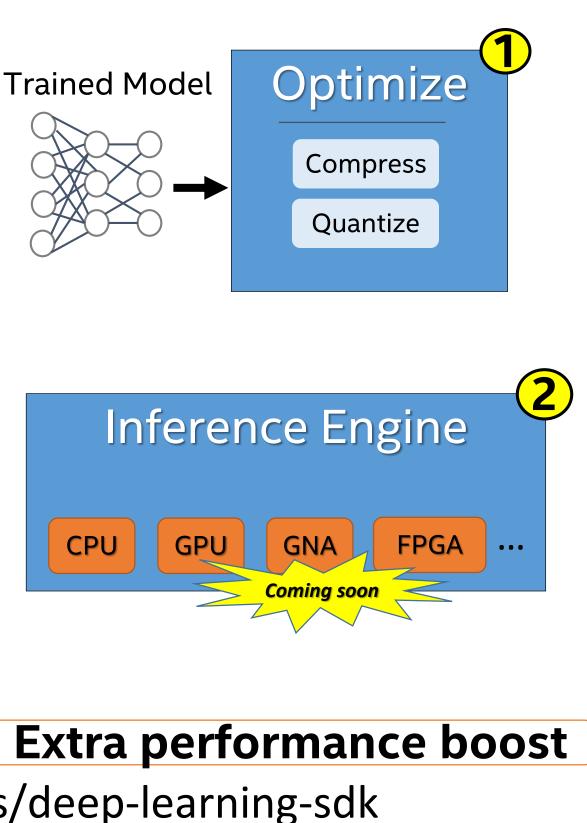




Intel DL SDK, GNA Native Library

2		
it cannot be swapped out)		
nning)		
layers of network		
st completes or timeout		

acting HW from developers



YSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions



What You Are Seeing: Live ASR Demonstration

Video playing TED talk by Aimee Mullens: The opportunity of adversity Customer speech recognizer performing local real-time transcription Acoustic model: 6-layer DNN w/ 2048 hidden nodes trained on 3000 hrs Acoustic likelihood scoring is selectable:

• Native (customer optimized), GNA SW emulation (CPU), or GNA (HW) Performance monitor shows drop in CPU residency with GNA HW Corresponding drop in power consumption not shown

Video playback with realtime transcription.

Recognition results match closed captioning.

Offload saves about 50% of a 1.1GHz Atom[™] core in this configuration*

s (e.g., higher CPU clock, use of more CPU cores, etc.) may improve performance at cost of higher power consumption

What You Are Seeing: Performance Demonstration

- DNN: additional layer added mapping outputs to phones
- Ordered by class: non-speech, unvoiced fricatives, voiced stops, unvoiced stops, voiced fricatives, liquids & glides, nasals, front vowels, mid vowels, back vowels, diphthongs
- LSTM: CTC-trained phone outputs in natural order

Batch scoring of 7-layer 2048 hidden node DNN

Log-likelihoods color coded

Low likelihood

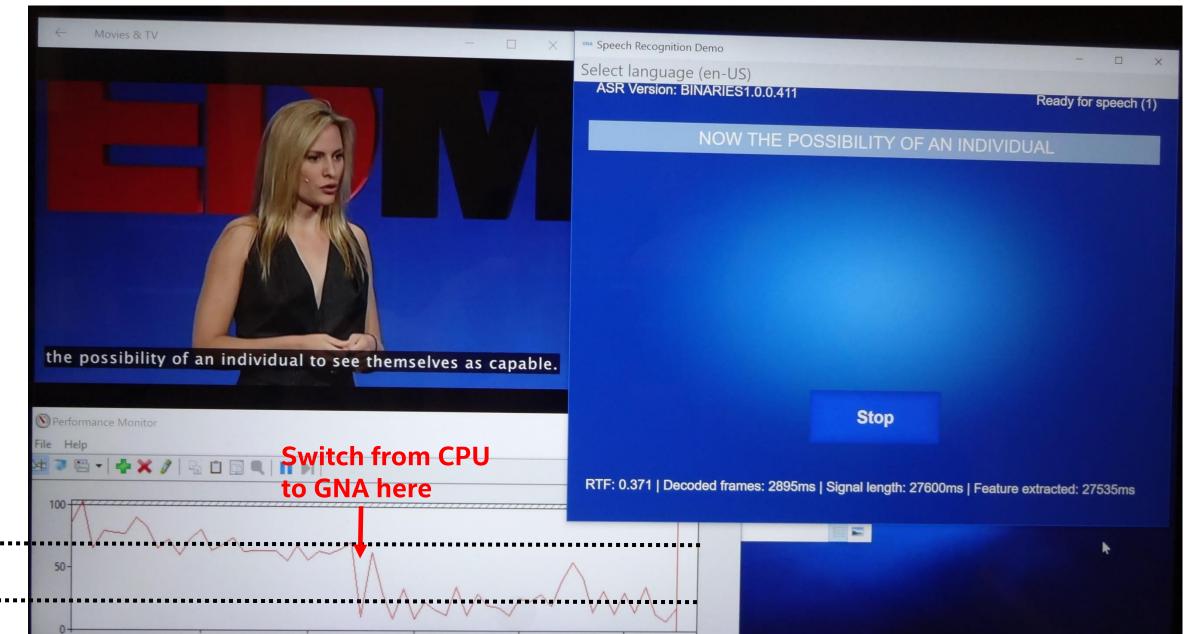
High likelihood

CPU utilization in Perfmon

Scoring on GNA is 3x faster than 1.1GHz Atom CPU in this configuration**

> **Note that other configurations (e.g., higher CPU clock, etc.) may have lower utilization benefit while retaining significant power reduction Intel® Quark™, Intel Atom®, Intel® Core™, and Intel® Deep Learning SDK are a trademark of Intel Corporation or its subsidiaries in the U.S. and/or other countries For more information on benchmarking go to www.intel.com/benchmark

Andrzej Polonski andrzej.polonski@intel.com Intel Technology Poland Gdansk, Poland



Real-time visualization of acoustic log-likelihoods

Repeating cycle: score on CPU, score on GNA, ...

Shows difference in scoring speed between CPU and GNA HW

WW Output Visualization	
	vowels fricatives non-speech
Control Panel – O X	Nerformance Monitor
Utterance length: 20 seconds	M > M + A / B C M A H N
Previous run on CPU: 20.721 Current run on CPU:	100-
Previous run on HW: 6.586 Current run on HW: 3.7	
SW/HW gain: 3.15x DNN grouping: 8 -	Running on an
	on and on and on and
Start DNN Start LSTM Stop	CPU CPU CPU
	9:54:29 AM 9:55:01 AM 9:55:31 AM 9:56:01 AM
demo_pack	5, 94, 63 MIVI
	Last Average Minimum Maximum Duration
	Show Color Scale Counter Instance Parent Object Co
GNADemo	Image: Text and the second s
	5