Computing in Memory: Speech Recognition

Spansion, Nuance, and Objective Analysis
Session 302-B

Introduction

- Industrial design and UI are driving adoption of new technology products
- Natural HMIs involve extremely complex systems
- Speech is the latest HMI being adopted

What are the challenges and solutions being explored in improving speech recognition and Natural Language Understanding?

Thursday, August 23, 12

Industry Experts

- Gabi Artzi, Solutions Architect, Automotive Business Unit, Nuance

- Rich Fastow, Director, Device Technology Engineering, Spansion

- Jim Handy, Director and Chief Analyst, Objective Analysis

Moderator: Stephan Rosner, VP, Software and Systems, Spansion

Thursday, August 23, 12

Flash Memory Summit
Session 302-B
Computing in Memory: Speech Recognition

Gabi Artzi, Solutions Architect

August 2012

Nuance is Innovating in Major Consumer Markets
Cars, Phones, eReaders, GPS, Computers, Copiers, …

AUTOMOTIVE SOLUTIONS

NUANCE

Thursday, August 23, 12

A Tipping Point?

Developer Community

Computational Resources
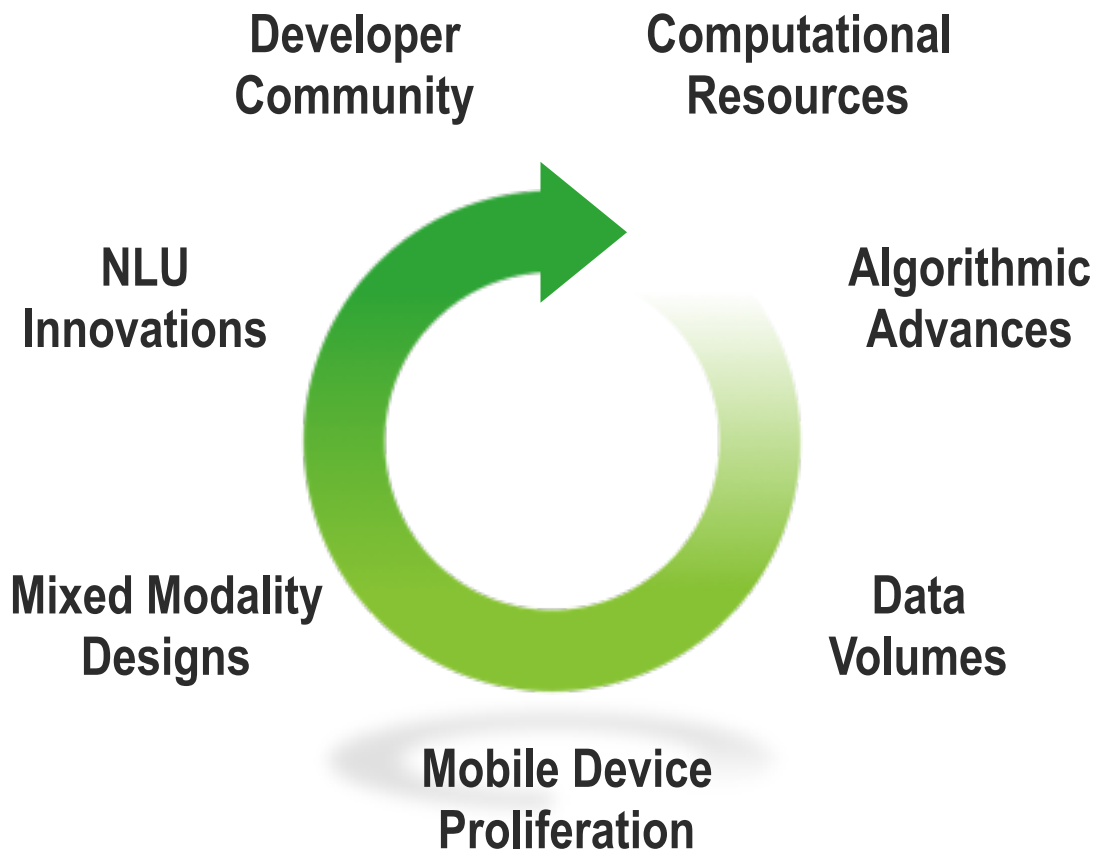
NLU Innovations

Algorithmic Advances

Mixed Modality Designs

Data Volumes

Mobile Device Proliferation

- 2 billion mobile transactions annually
- 6 billion handsets deployed
- 70 million cars shipped
- 10 thousand mobile app developers
- 500 thousand physicians
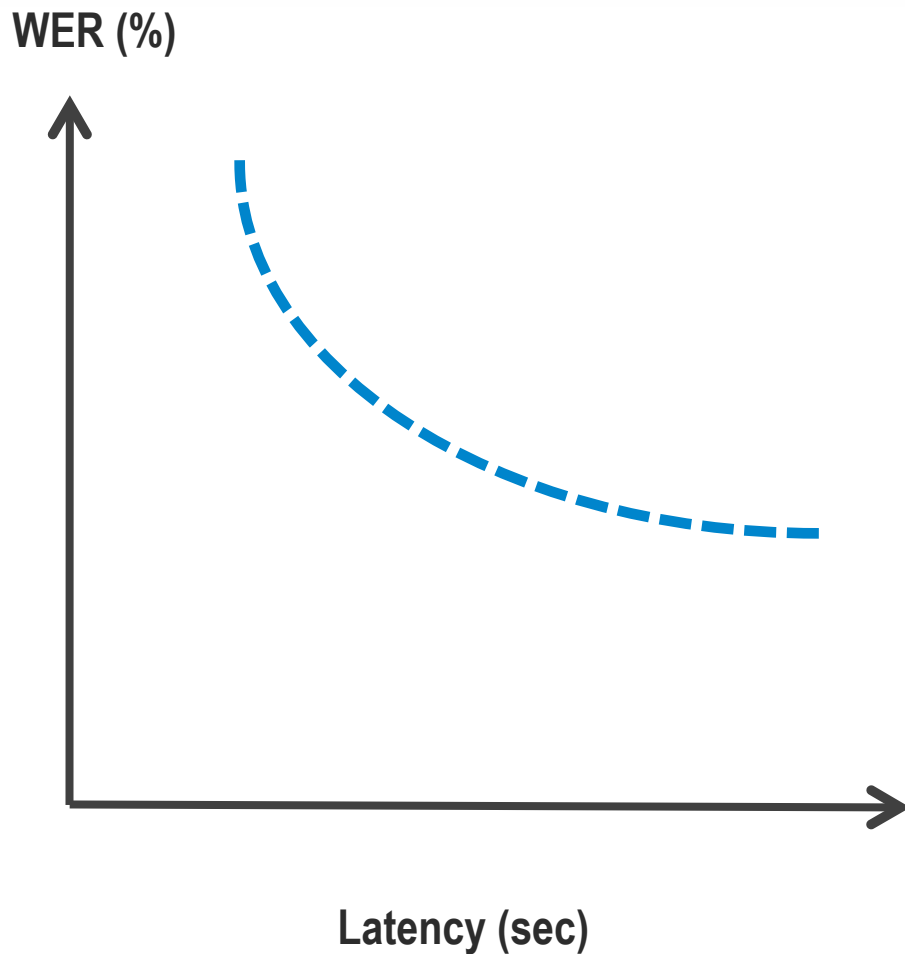- 51 languages supported

NUANCE

A Tipping Point?

Developer Community

Computational Resources

NLU Innovations

Algorithmic Advances

Mixed Modality Designs

Data Volumes

Mobile Device Proliferation

- 2 billion mobile transactions annually
- 6 billion handsets deployed
- 70 million cars shipped
- 10 thousand mobile app developers
- 500 thousand physicians
- 51 languages supported

NUANCE

Why an Acoustic Coprocessor?

- Offload CPU

- Increase Acoustic Models Size
    - Drives Higher Recognition Accuracy

- Reduce RAM consumption

- Reduce Speech Recognition Latency
    - Drives Better User Experience

NUANCE

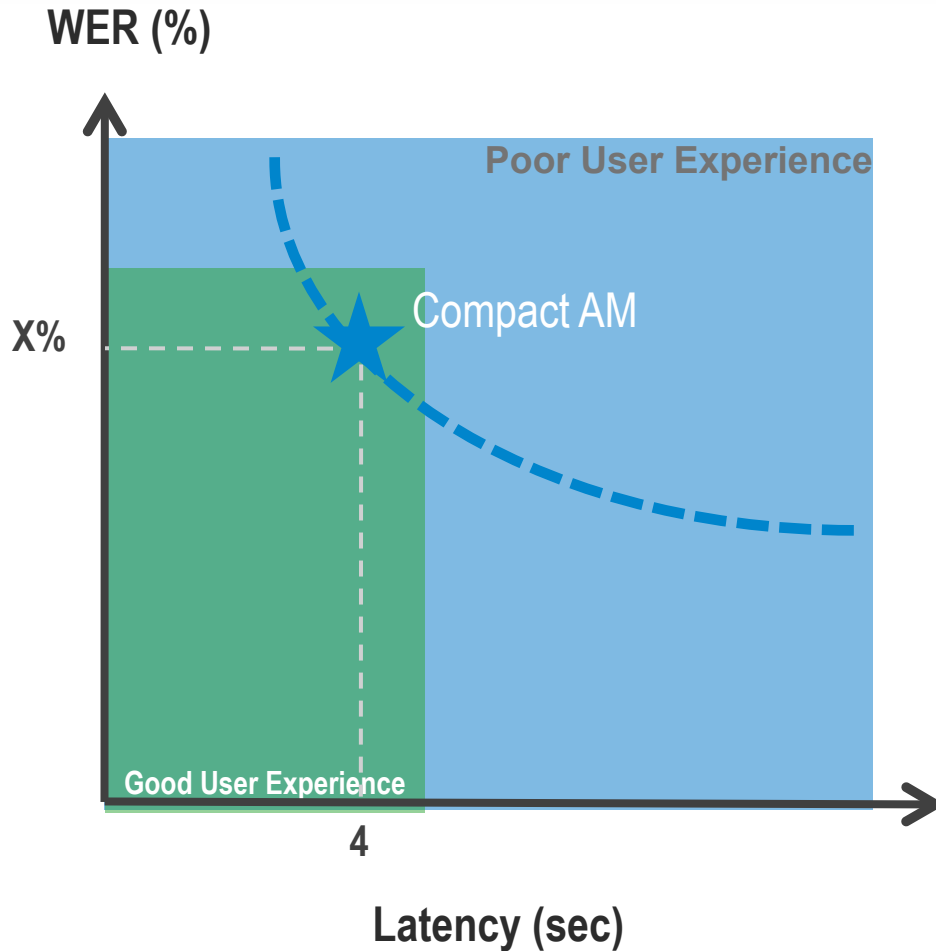# Speech Recognition Quality Metrics & User Experience
(Experiments done by Spansion)

WER (%)

Latency (sec)

- Quality of speech recognition solutions is determined by
  - *Accuracy* = Word Error Rate (WER)
  - *Latency* = system responsiveness
- Today's Systems use "Compact" models
  - Optimized for embedded systems
  - Trade-off between Accuracy and latency
- Next-Generation "Full" models
  - Needed for next-generation applications
  - 15-20% reduction in error rate
  - Unacceptable latencies

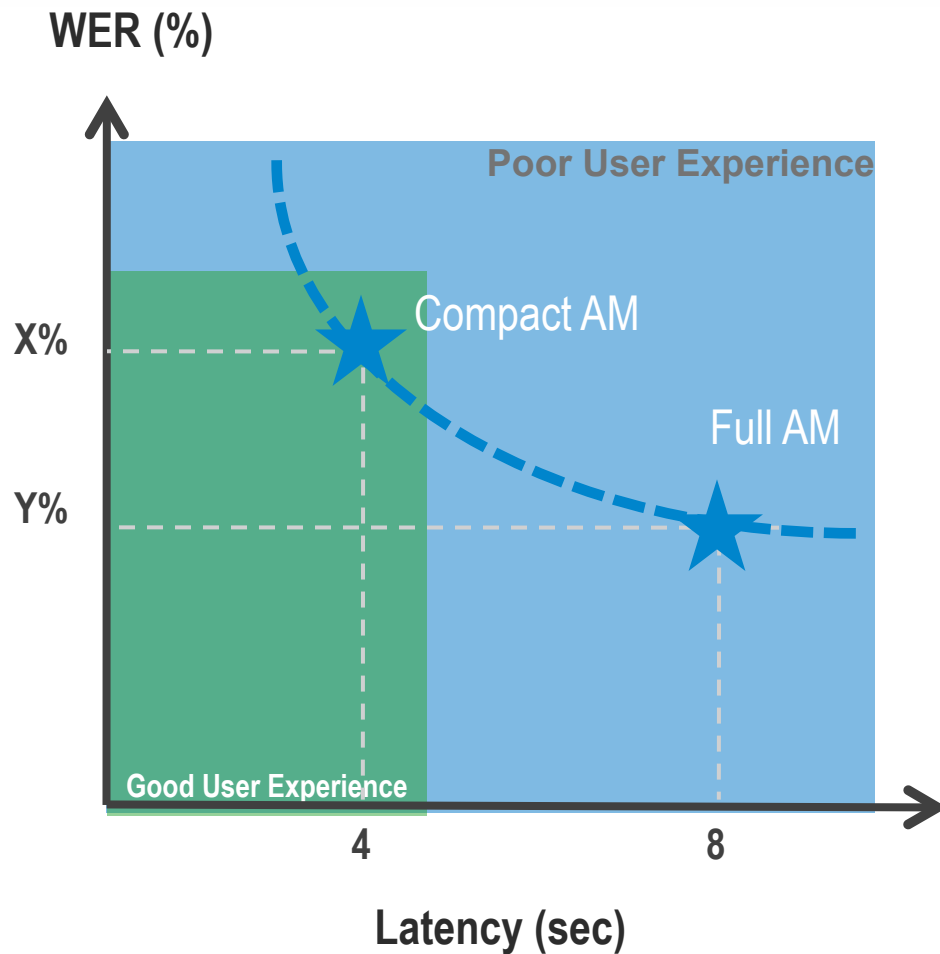# Speech Recognition Quality Metrics & User Experience
(Experiments done by Spansion)

WER (%)

Poor User Experience

Compact AM

X%

Good User Experience

4

Latency (sec)

- Quality of speech recognition solutions is determined by
    - *Accuracy* = Word Error Rate (WER)
    - *Latency* = system responsiveness
- Today's Systems use "Compact" models
    - Optimized for embedded systems
    - Trade-off between Accuracy and latency
- Next-Generation "Full" models
    - Needed for next-generation applications
    - 15-20% reduction in error rate
    - Unacceptable latencies

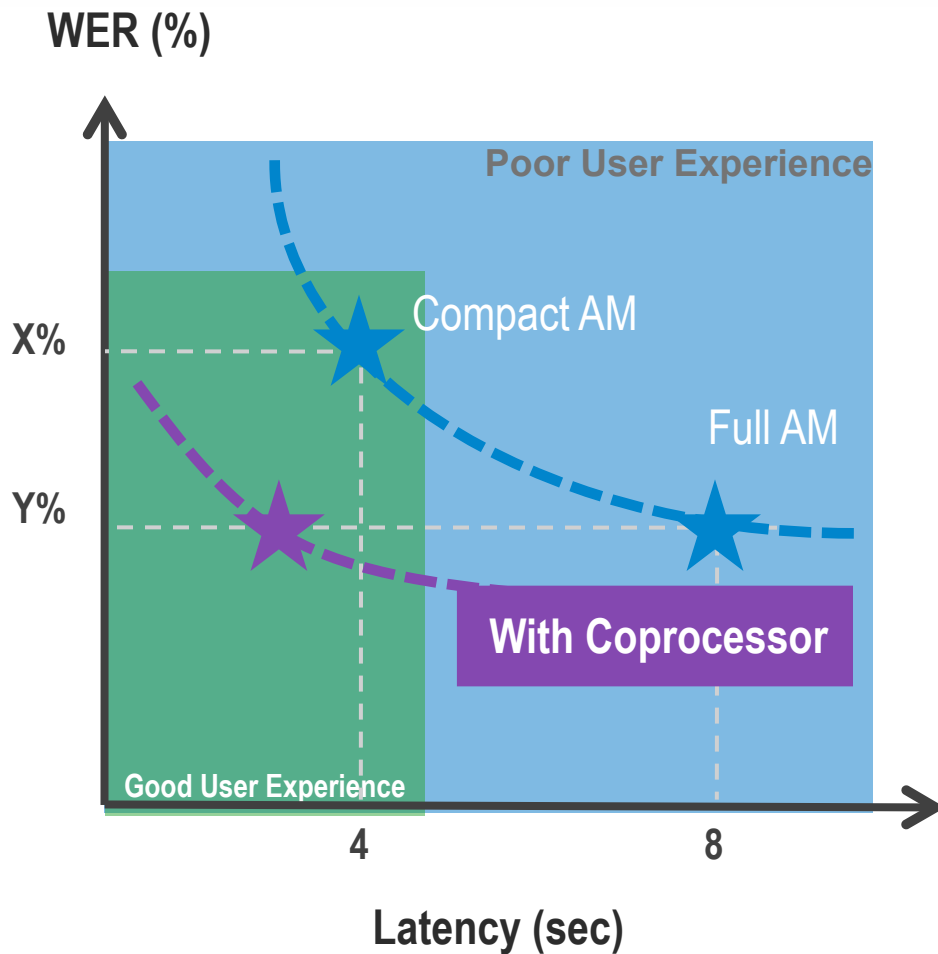# Speech Recognition Quality Metrics & User Experience
(Experiments done by Spansion)

WER (%)

Latency (sec)

- Quality of speech recognition solutions is determined by
  - *Accuracy* = Word Error Rate (WER)
  - *Latency* = system responsiveness
- Today's Systems use "Compact" models
  - Optimized for embedded systems
  - Trade-off between Accuracy and latency
- Next-Generation "Full" models
  - Needed for next-generation applications
  - 15-20% reduction in error rate
  - Unacceptable latencies

SPANSION®   NUANCE

Speech Recognition Quality Metrics & User Experience
(Experiments done by Spansion)

WER (%)

Poor User Experience

Compact AM

Full AM

X%

Y%

With Coprocessor

Good User Experience

4

8

Latency (sec)

- Quality of speech recognition solutions is determined by
  - *Accuracy* = Word Error Rate (WER)
  - *Latency* = system responsiveness
- Today's Systems use "Compact" models
  - Optimized for embedded systems
  - Trade-off between Accuracy and latency
- Next-Generation "Full" models
  - Needed for next-generation applications
  - 15-20% reduction in error rate
  - Unacceptable latencies

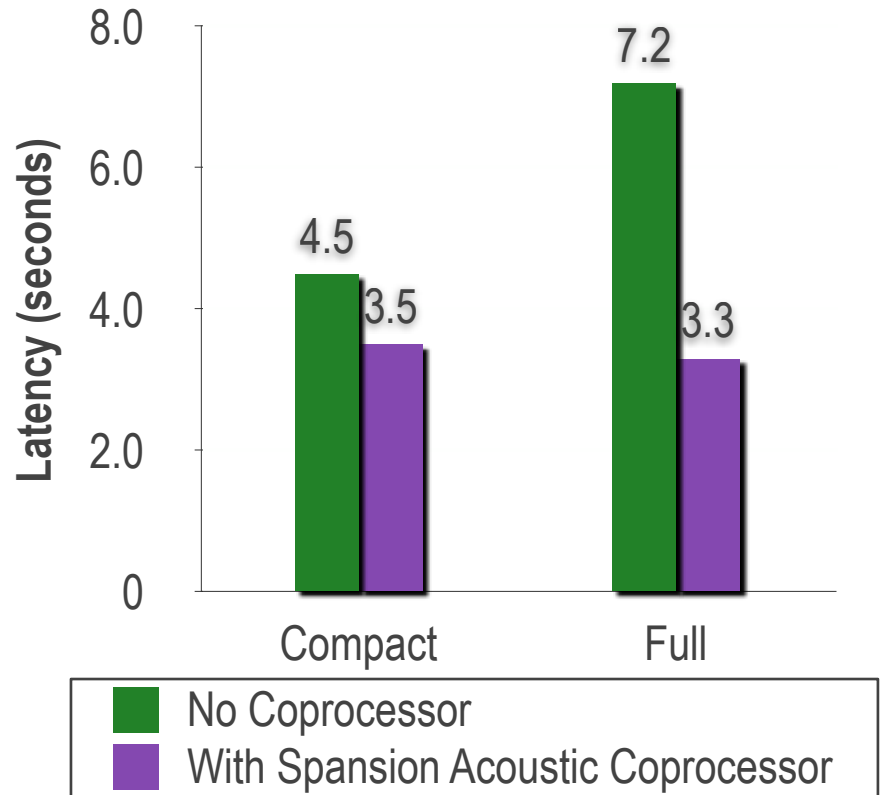**Spansion Acoustic Coprocessor enables optimal use of Full acoustic models**

SPANSION®   NUANCE®

Thursday, August 23, 12

Benchmark results (ENU and GDE 1-shot VDE)
(Experiments done by Spansion)
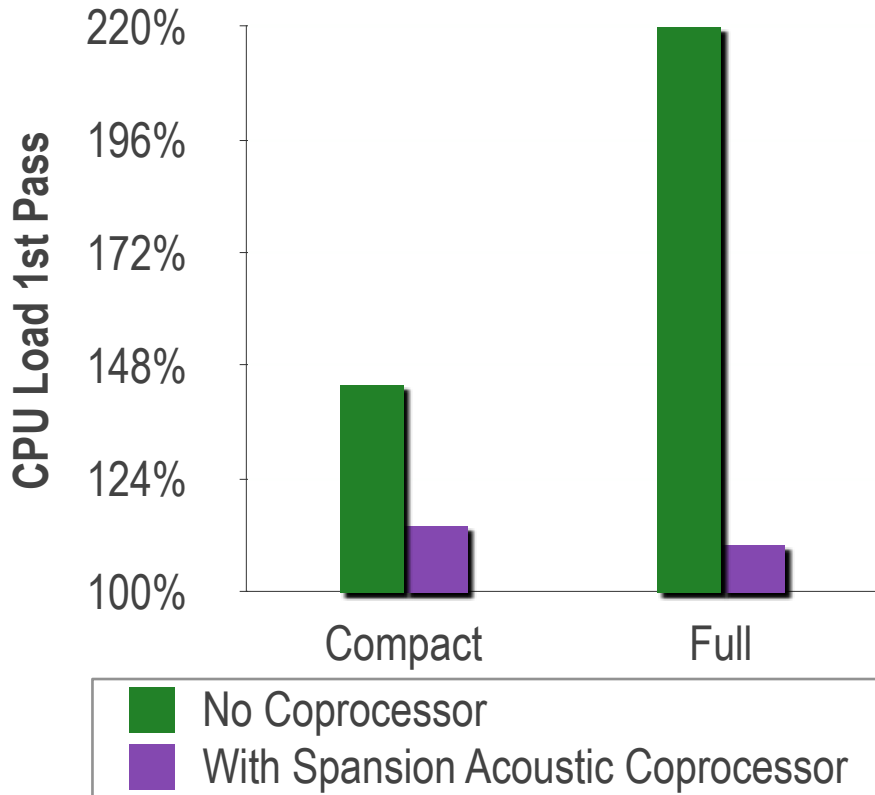
CPU gain:
- ➢>20% for Compact models
- ➢>40% for Full models

Latency reduction:
- ➢>25% for Compact models
- ➢>50% for Full models

Thursday, August 23, 12

Flash Memory Summit

Computing in Memory: Speech Recognition
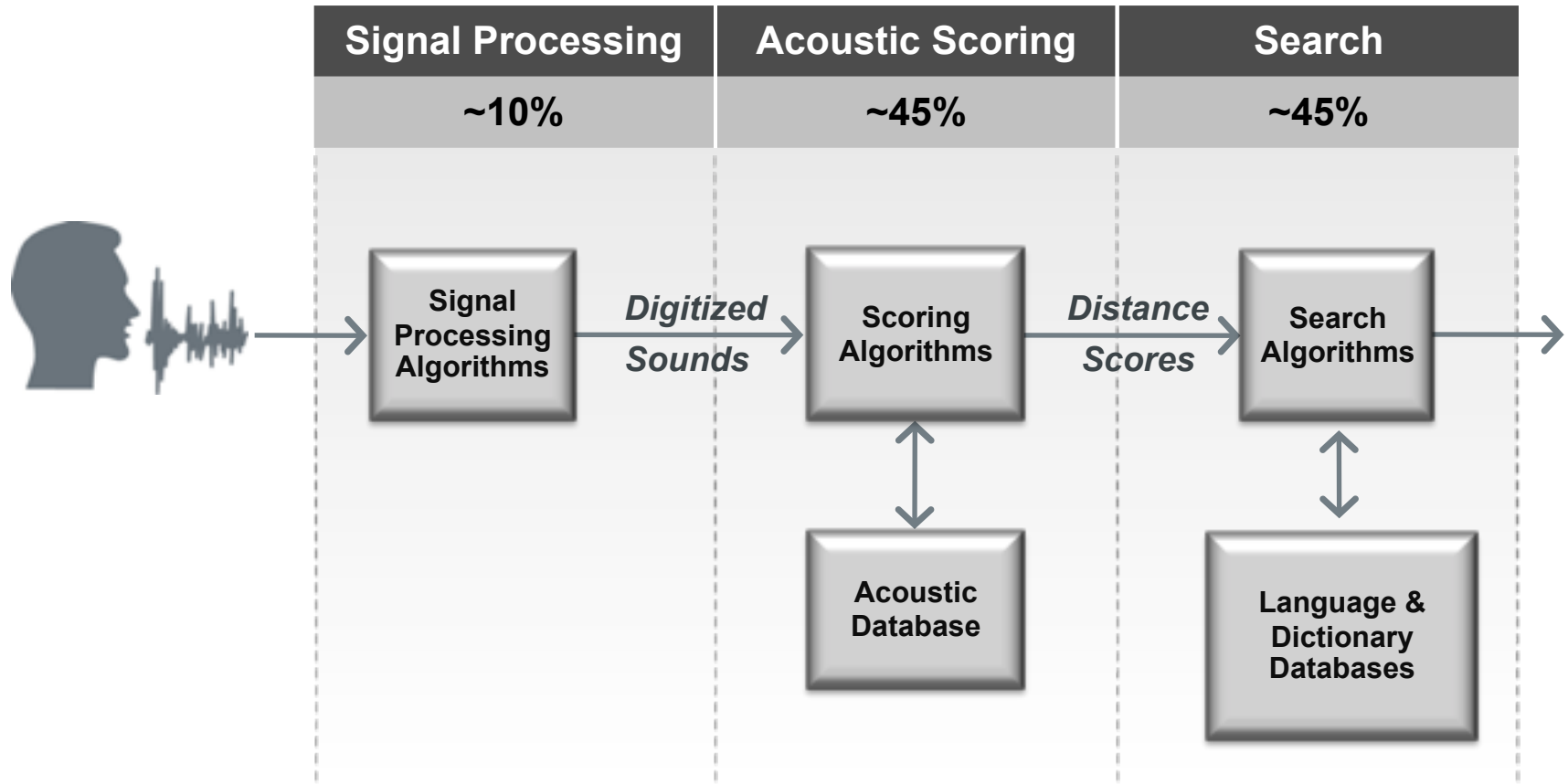Spansion® Acoustic Coprocessor

Rich Fastow
August 23, 2012

Thursday, August 23, 12

Spansion Acoustic Coprocessor
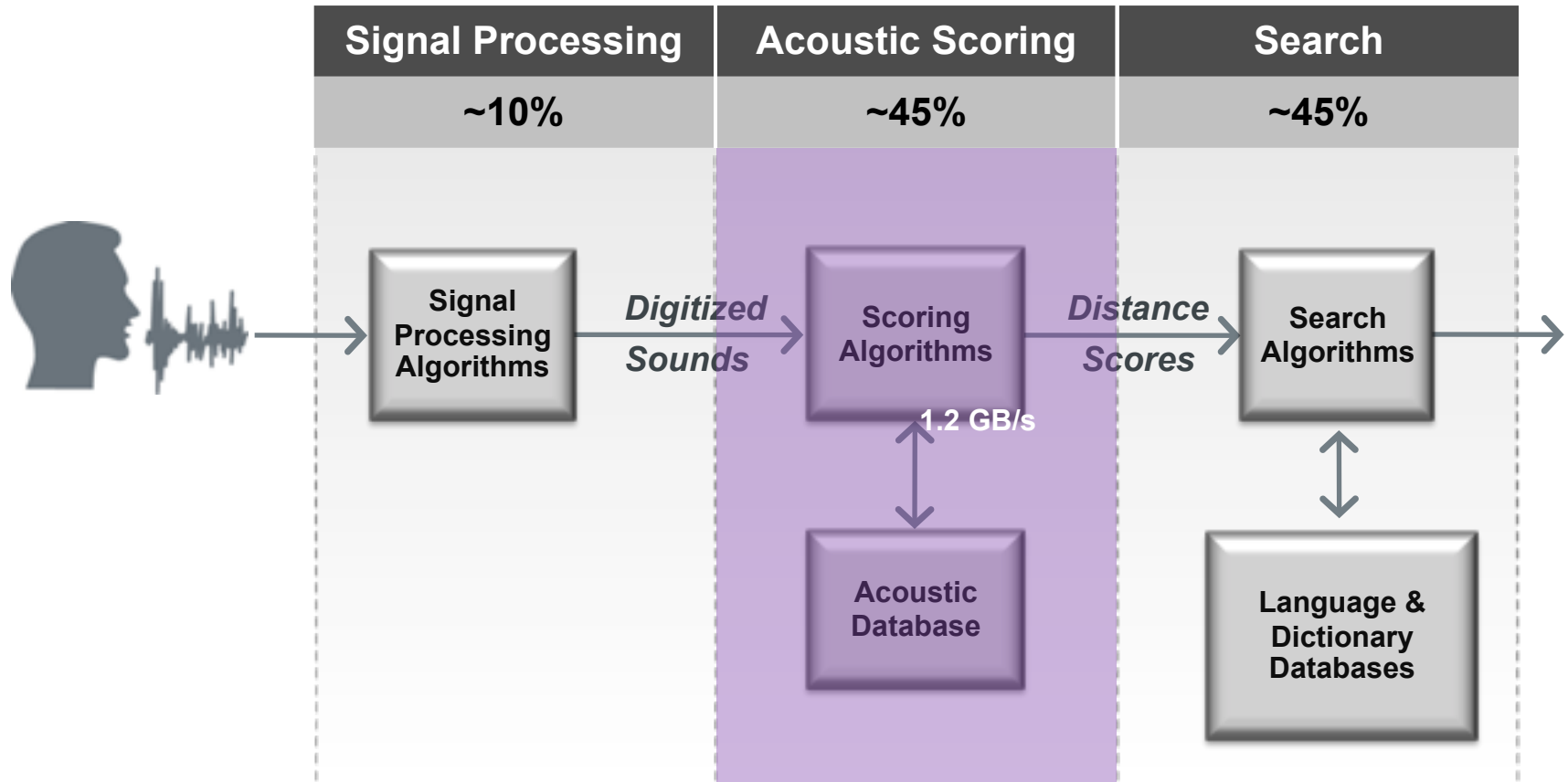

Spansion® Acoustic Coprocessor

Combines custom logic and high-speed NOR flash memory to deliver faster, more accurate voice recognition

Enhances the user interface in embedded applications, creating a more compelling voice experience for consumers

SPANSION®

Thursday, August 23, 12

Voice Recognition – Custom <u>Logic + Memory</u> Solution

Voice Recognition – Custom <u>Logic + Memory</u> Solution

Signal Processing	Acoustic Scoring	Search
~10%	~45%	~45%

Signal Processing Algorithms → *Digitized Sounds* → Scoring Algorithms → *Distance Scores* → Search Algorithms
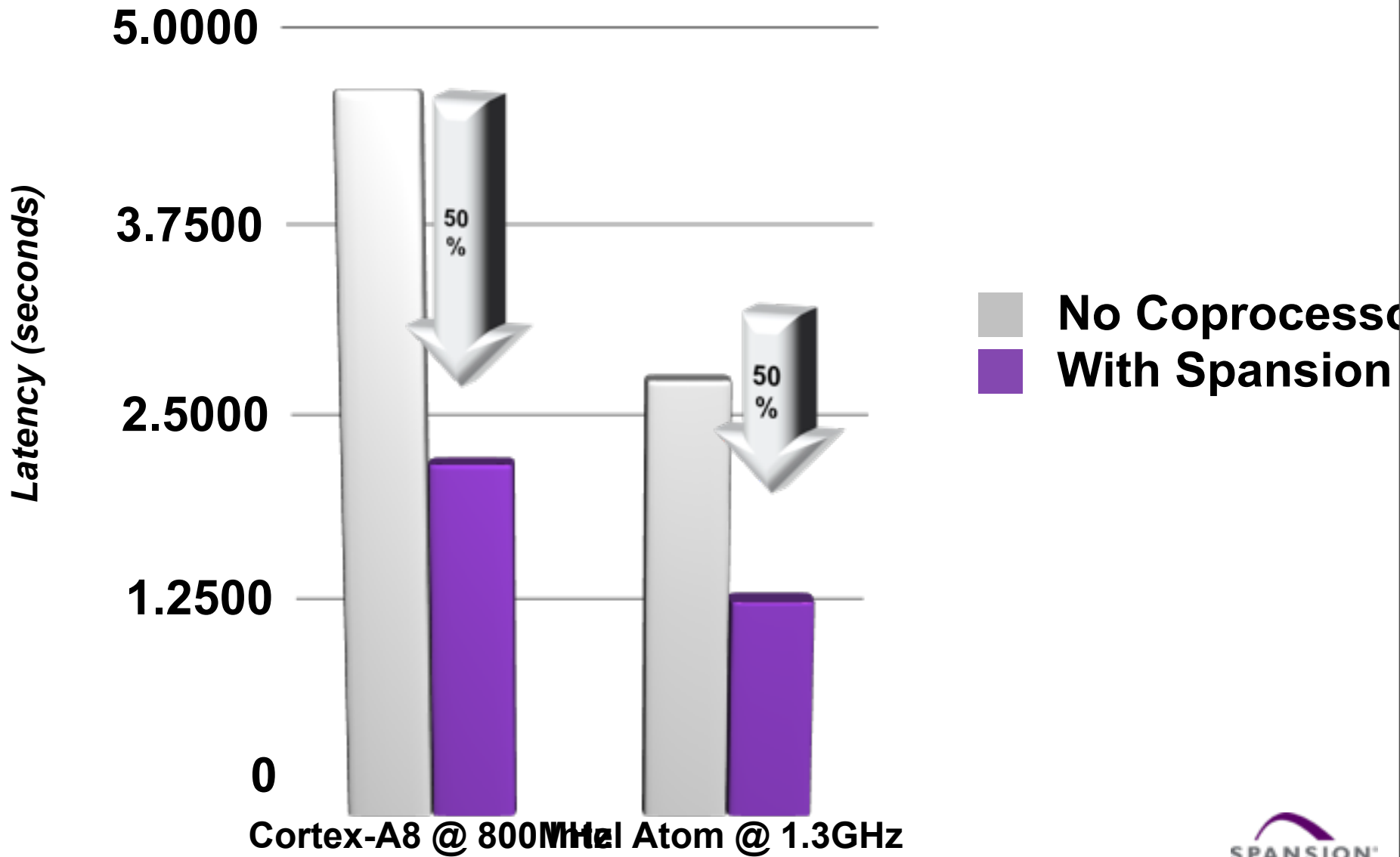
1.2 GB/s

Acoustic Database

Language & Dictionary Databases

Spansion Acoustic Coprocessor accelerates the speech process by combining custom logic and high speed on-chip NOR flash memory

SPANSION®

Spansion Acoustic Coprocessor Concept
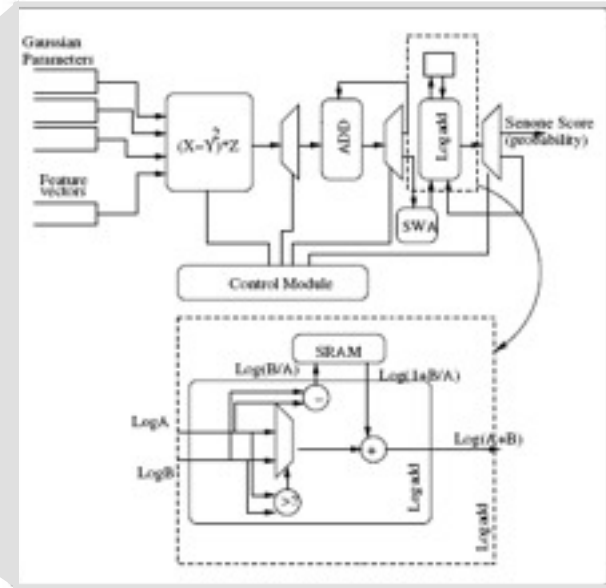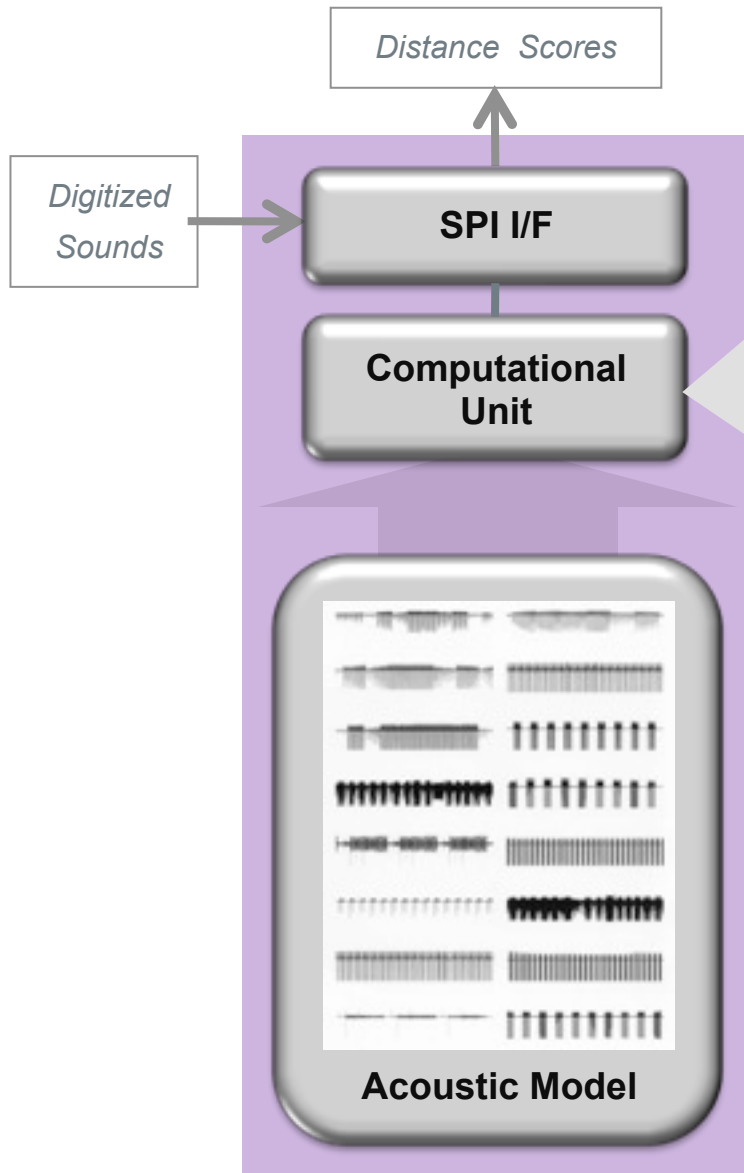
- **65nm NOR flash technology**

- **Computational Unit consists of 8 ALUs, customized to calculate the distance between the incoming sound and the Acoustic Database**

- **Acoustic Database stored in on-chip flash**

- **High BW, on-chip data bus**

- **Low BW, SPI host interface**

© 2012 Spansion, Inc.

Thursday, August 23, 12

Infotainment System Characterization (Latency)

© 2012 Spansion, Inc.
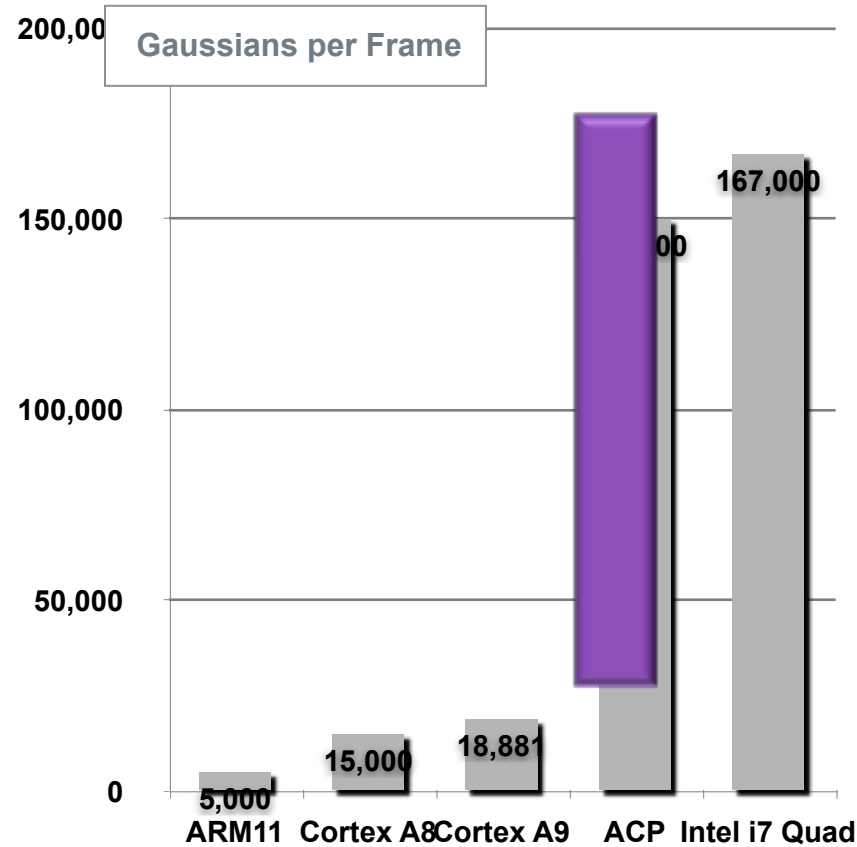
Thursday, August 23, 12

The Future of Embedded Voice Recognition

Spansion Acoustic Coprocessor:

- Improves Voice Recognition Response Time
  - Up to 50% latency reduction with today's acoustic databases

- Enhances Overall System Performance
  - Up to 50% CPU load reduction

- Enables an Improved Voice Recognition Experience
  - Natural Language Understanding (NLU)
  - Multi-lingual /multi-accent support

SPANSION®

Thursday, August 23, 12

Spansion Acoustic Coprocessor - Performance Benchmark

Gaussians per Frame

Chart values: ARM11 5,000; Cortex A8 15,000; Cortex A9 18,881; ACP; Intel i7 Quad 167,000

Executing Nuance Acoustic Model scoring algorithms on single processors versus implementing in dedicated HW in Spansion Acoustic Coprocessor

SPANSION

Thursday, August 23, 12

www.spansion.com

Thursday, August 23, 12

Why Speech is Big

Jim Handy

OBJECTIVE ANALYSIS

Moore's Law Shrinks Chips
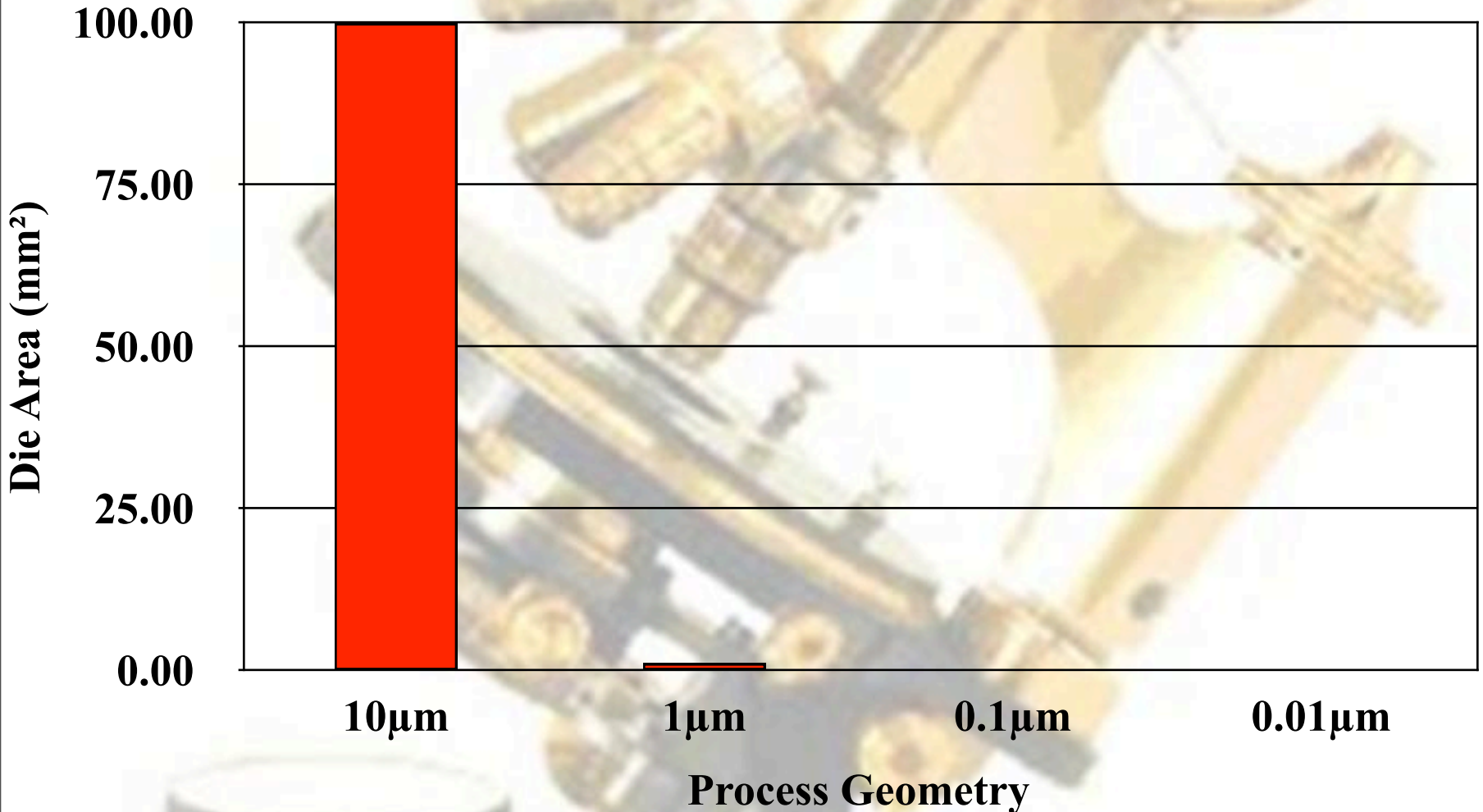
Die Area (mm²) vs **Process Geometry**

100.00 — 75.00 — 50.00 — 25.00 — 0.00

10μm   1μm   0.1μm   0.01μm

Thursday, August 23, 12

I/O Limits Product Size

…and Usability

Visual Interfaces Distract Drivers

Thursday, August 23, 12

Speech: The Smallest UI

Thursday, August 23, 12

Many Potential Speech Markets

- 2.5 trillion voicemails
- 2 billion mobile transactions annually
- 6 billion handsets
- 300 million computers
- 70 million automobiles
- 10 thousand mobile app developers
- 500 thousand physicians

OBJECTIVE ANALYSIS – www.OBJECTIVE-ANALYSIS.com

Many Potential Speech Markets

- 2.5 trillion voicemails
- 2 billion mobile transactions annually
- 6 billion handsets
- 300 million computers
- 70 million automobiles
- 10 thousand mobile app developers
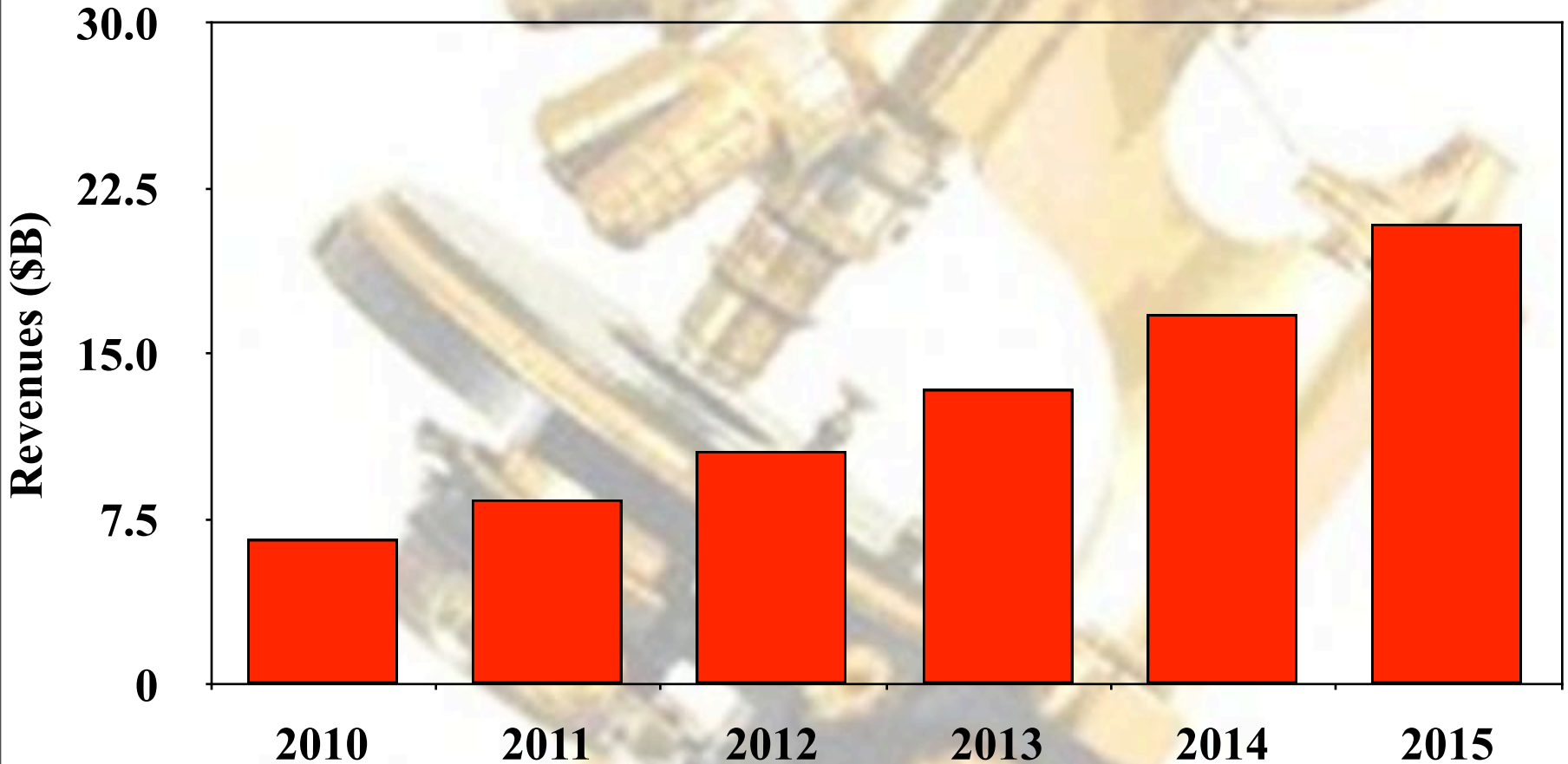- 500 thousand physicians

OBJECTIVE ANALYSIS – www.OBJECTIVE-ANALYSIS.com

Thursday, August 23, 12

Result: Speech Interface Growth

Source; GIA 2010

OBJECTIVE ANALYSIS – www.OBJECTIVE-ANALYSIS.com

Thank You!

Jim Handy

OBJECTIVE
ANALYSIS

Session Q&A

Thursday, August 23, 12