

# SAMOS'16 - Post talk notes

- **Contacting ARM – Please include SAMOS in title**
  - Potential Arch Research (Stuart's list) - Stuart Biles <[Stuart.Biles@arm.com](mailto:Stuart.Biles@arm.com)>
    - Also accelerators / Heterogeneous Compute
  - Potential Memory Research (Stephan's list) - Stephan Diestelhorst <[Stephan.Diestelhorst@arm.com](mailto:Stephan.Diestelhorst@arm.com)>
    - Also GEM5
  - Security and Specifications (Alastair's list) - Alastair Reid <[Alastair.Reid@arm.com](mailto:Alastair.Reid@arm.com)>
  - Participation in EC-funded Collaborative Projects - [emre.ozar@arm.com](mailto:emre.ozar@arm.com)
  - Research Partnerships – Chris Doran <[Chris.Doran@ARM.com](mailto:Chris.Doran@ARM.com)>
  - General Qs / IP Requests - [Eric.Hennenhoefer@ARM.com](mailto:Eric.Hennenhoefer@ARM.com)
    - New website in the works to make IP requests easier
- **ARM Research Summit – Sept 15<sup>th</sup>-16<sup>th</sup> Cambridge UK**
  - <https://developer.arm.com/research/summit>

# Standard Disclaimer

Everything that follows is the personal opinion of the speaker.

# ARM Research

**ARM**

Eric Hennenhoefer <[eric.hennenhoefer@ARM.com](mailto:eric.hennenhoefer@ARM.com)>  
VP Research

SAMOS'16

**IBM**  
*PowerPC™*



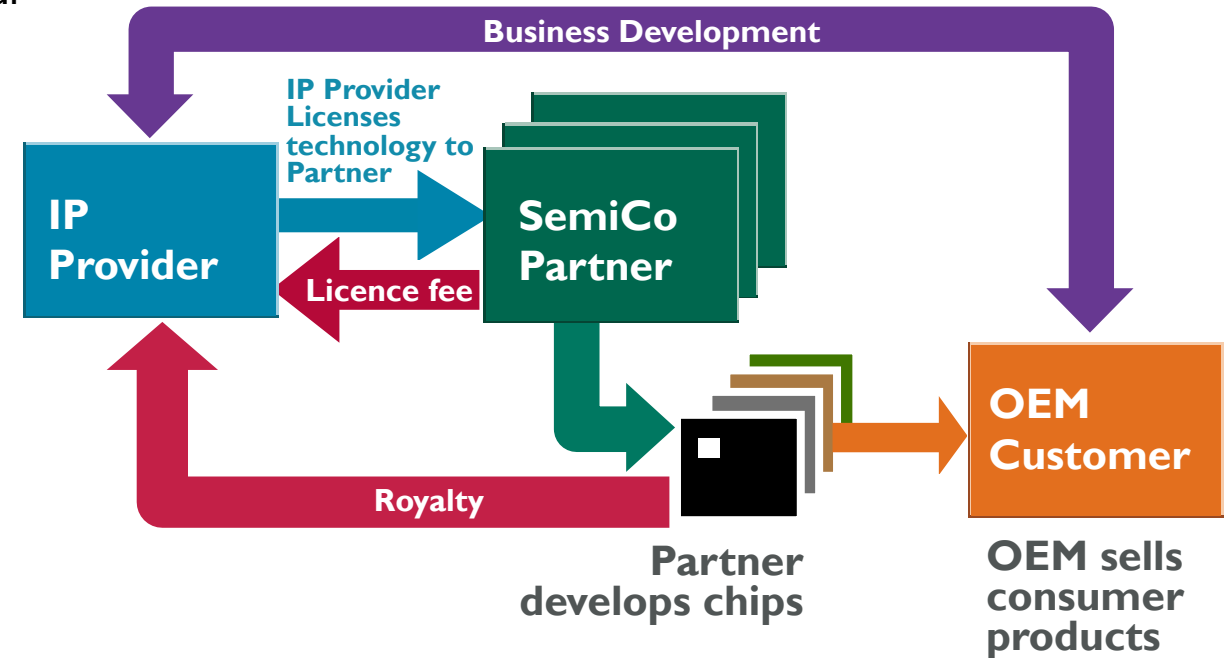


# Agenda

- ARM Overview
- Why is Eric at SAMOS?
- Is Computer Architecture Dying?
- About that Funnel... Yale's wrong, the future is commodity SI
- Research Enablement (free stuff)
- Parting thoughts for students

# A Partnership Business Model

- Have a viable business model
  - Understand how everyone in the value chain can be successful
  - Long term sustainability
- Design once and reuse is fundamental
  - Spread the cost amongst many partners
  - Technology reused across multiple applications
  - Creates market for ecosystem to target
    - Re-use is also fundamental to the ecosystem
- Upfront license fee
  - Covers the development cost
- Ongoing royalties
  - Typically based on a percentage of chip price
  - Vested interest in success of customers



Approximately **1250** licenses  
Grows by **~100** every year

More than **350** potential  
royalty payers

**12bn+** ARM-powered chips in 2014  
**>25%** CAGR over last 5 years

# ARM Technology

Advanced consumer products are incorporating more and more ARM technology – from processor and multimedia IP to software

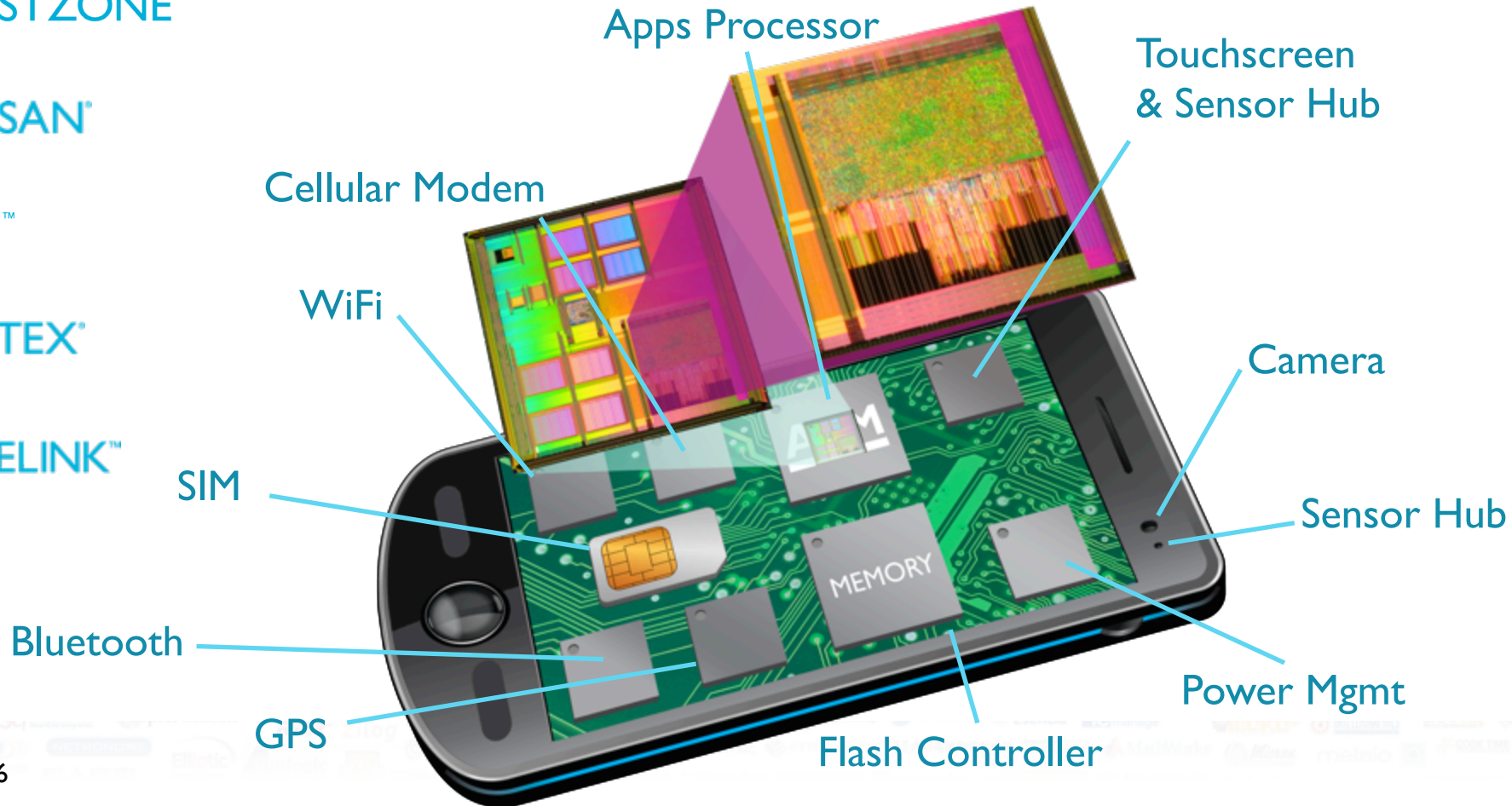
**ARM TRUSTZONE**  
System Security

**ARM ARTISAN**  
Physical IP








**ARM MALI**  
Visual Technology

**ARM CORTEX**  
Processor Technology

**ARM CORELINK**  
Processor System IP



# Tiny Sensors to Industrial Grade Compute

Mixed Signal IoT Sensors	Embedded, Wearables	Communications and Control	Smartphone ,Tablet, IVI, ADAS	Infrastructure and Servers
<2mm <sup>2</sup> 	2-8mm <sup>2</sup> 	25-40mm <sup>2</sup> 	25-40mm <sup>2</sup>  50-80mm <sup>2</sup> 	25-100mm <sup>2</sup>  ↔ 80-100mm <sup>2</sup> 
<10uW	<10mW	10s of mW	100s of mW	1-5W <span style="margin-left: 20px;">&gt;5W</span>

RTOS                      RTOS, Simple UI                      Rich OS, Sophisticated UI                      Enterprise Software

Under 50c to over 25\$

Diverse range of markets, each with its own unique demands and opportunities



# Goal – Create a Diverse and Vibrant Ecosystem

The image displays a large collection of logos for various partners, organized into three main categories:

- Software, Training and Consortia Partners:** This section includes logos for companies like ACUSTIC, ACCESS, AKAE, ANACOM, and many others, representing a wide range of software and training providers.
- Silicon Partners:** This section features logos for semiconductor and silicon-related companies such as ARM, AMD, Intel, and various regional chip manufacturers.
- Design Support Partners:** This section lists logos for design support and tool providers, including companies like ANSYS, Cadence, and various design houses.

The logos are densely packed and cover the entire central area of the slide, illustrating a diverse and vibrant ecosystem.



# Why is Eric at SAMOS?

ARM Research is the primary stakeholder for ARM / academia relationship

Because...

ARM Research decided that stronger academic ties will make the group more successful..

- **Untested Assumption:**
  - ARM is uniquely positioned to extend the ARM partnership model to create a win/win academic research ecosystem which can accelerate technology wherever computing happens.
- **Goals of SAMOS keynote**
  - Communicate new strategy and activities
  - Test degree commonality between academic and ARM Research's goals

# Research Project Flow

## Mission

To create and transfer research knowledge across ARM impacting ARM's products and success in new markets.

## Strategic Objective #1

Build pipeline to create and bring future technology into ARM products

Technology

Explore

Refine

Tech Transfer

Pre-Competitive

Competitive

Pre-Competitive generates Knowledge

Competitive generates Value



# Introducing ARM Research

- About ARM Research
  - 3 – 7 years ahead of product teams
  - From advanced development to blue sky
  - Locations in Austin, Cambridge UK, San Jose, and Shanghai
- Objectives
  - Build a pipeline to create and bring future technology into ARM products
  - Create and maintain the technology roadmap
  - Enable academia and research partnerships

Limits of enablement: ARM Research can do almost anything, as long as we don't mess with the ongoing business...



# Research Focus Areas

## Memory & Interconnect



### Tracking and Driving Memory Roadmaps

- Leading future task group



### Going Beyond Evolutionary DRAM

- 3D stacked memories
- Intent-based interfaces



### NVM in the System

- Drive technology
- Ensure open standards

### Compute Near Memory

- Reduce data movement

## Architecture



### Embedded Efficiency

- TrustZone-M
- Improve code density and performance



### Security

- HW is the root of trust
- Make is easier to write secure SW



### Next Gen Arch

- Super secret stuff
- Use transistors more efficiently
- Accelerate key use cases



### New Apps

- Novel use cases

# Research Focus Areas

## Applied Silicon



### IoT Sensor Nodes

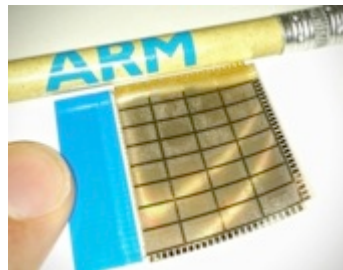
- Sub-threshold for 0.1x energy
- Energy optimized mixed-signal
- Extreme power gating

### Integrating everything

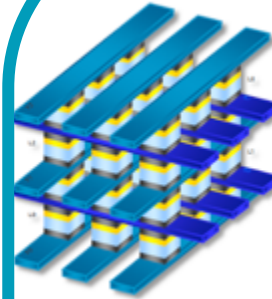
- Voltage regulators
- Energy harvesters
- Sensor interfaces

### Printed Electronics

- 1cent disposable MCUs
- Mapping the ecosystem

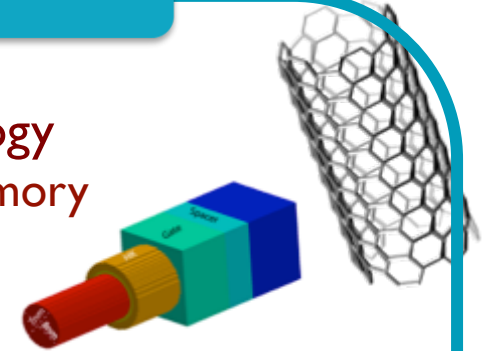


## Future Si Tech



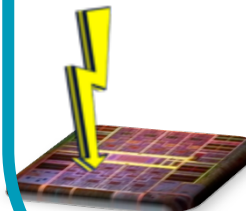
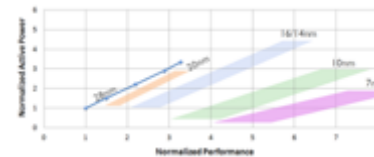
### Disruptive technology

- Next Big Thing Memory
- What's after MOS?
- 3DIC technology



### Predictive Technology Modeling

- Technology scaling entitlement
- Design-Technology Co-Optimization
- Next node device, patterning, ..



### Dependable Computing

- Detection, Correction, Security
- Robust power delivery



# Research Focus Areas

## Large Scale Systems



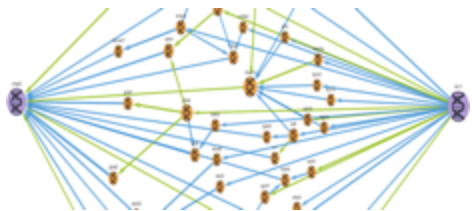
### High Performance Computing

- Enable the first ARM supercomputer



### sideARMs

- Compute near memory, network, and storage & standardize systems software interfaces



### Data Intensive

- Improving system efficiency for analytics workloads

## Design Integrity



### Formal Methods

- Formal Coherency Verification on Cortex®-A



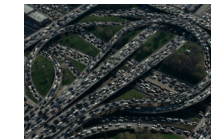
### Rain

- Deriving RTL checkers from Architecture specification



### CPU $\mu$ Arch Models

- Verifying implementations against executable spec



### Deadlock Dependency Models

- Design-time deadlock freedom for arbitrary interconnect topologies

# Research Focus Areas

## Emerging Applications



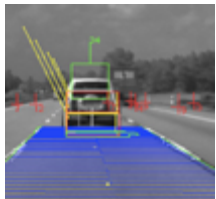
### Machine Learning

- Speech & image recog
- Neural networks



### Graphics Systems

- Full-system modeling
- System cache arch



### Computer Vision

- Emphasis on automotive
- Depth perception, object and motion tracking



### Mobile Systems

- Advanced workloads
- HW + SW system design
- Future devices

## Special Projects



### ARM motor

- Novel motor control

### Technology Roadmapping



### Technical Due Diligence



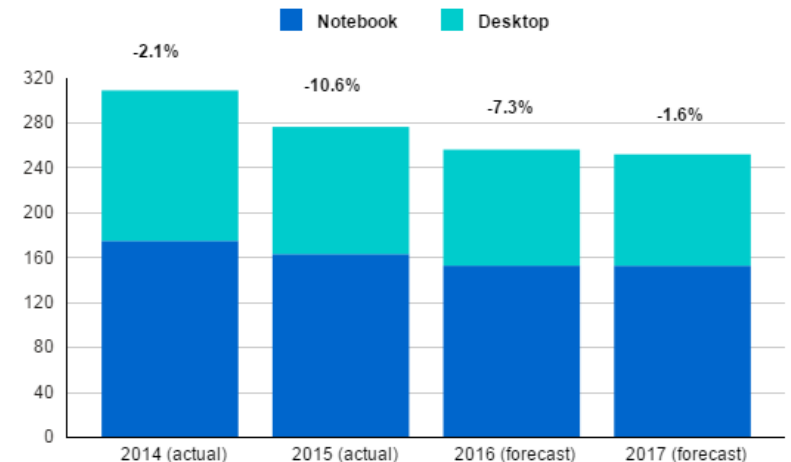
### Low Power Radio

# Is Computer Arch Dying? Is Innovation Slowing?

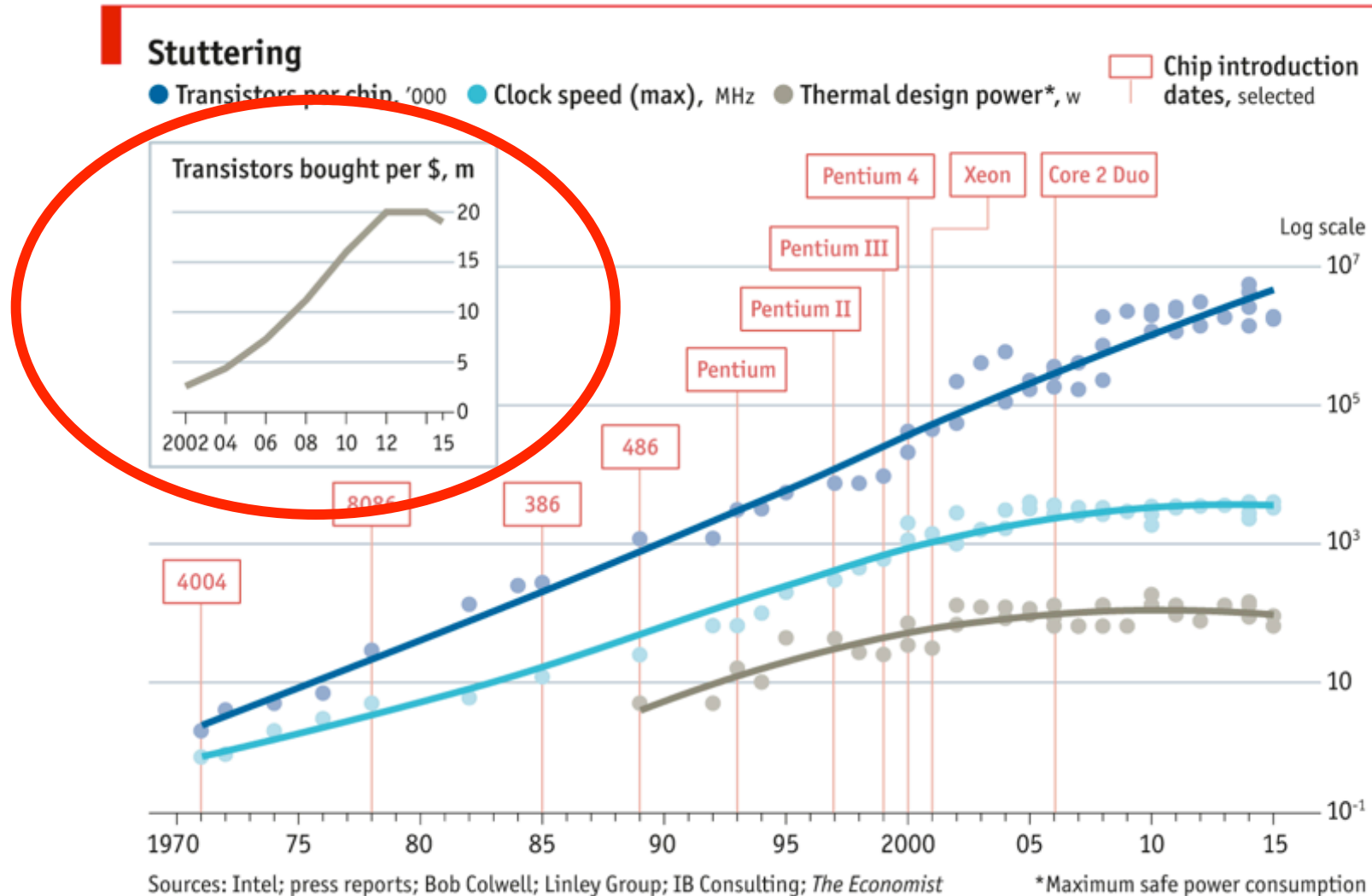
- Intel - Tick, Tock, Tock
- "peak smartphone"
  - Apple's Suppliers Projecting Weak Demand for iPhone 7 Due to 'Lack of Innovation'
  - Iphone 7 features— no audio jack or mute button, will 'Space Black' be the new 'Rose Gold'?



WW PC Forecast by Product Category, 2014-2017  
(Units in Millions and YoY Growth)

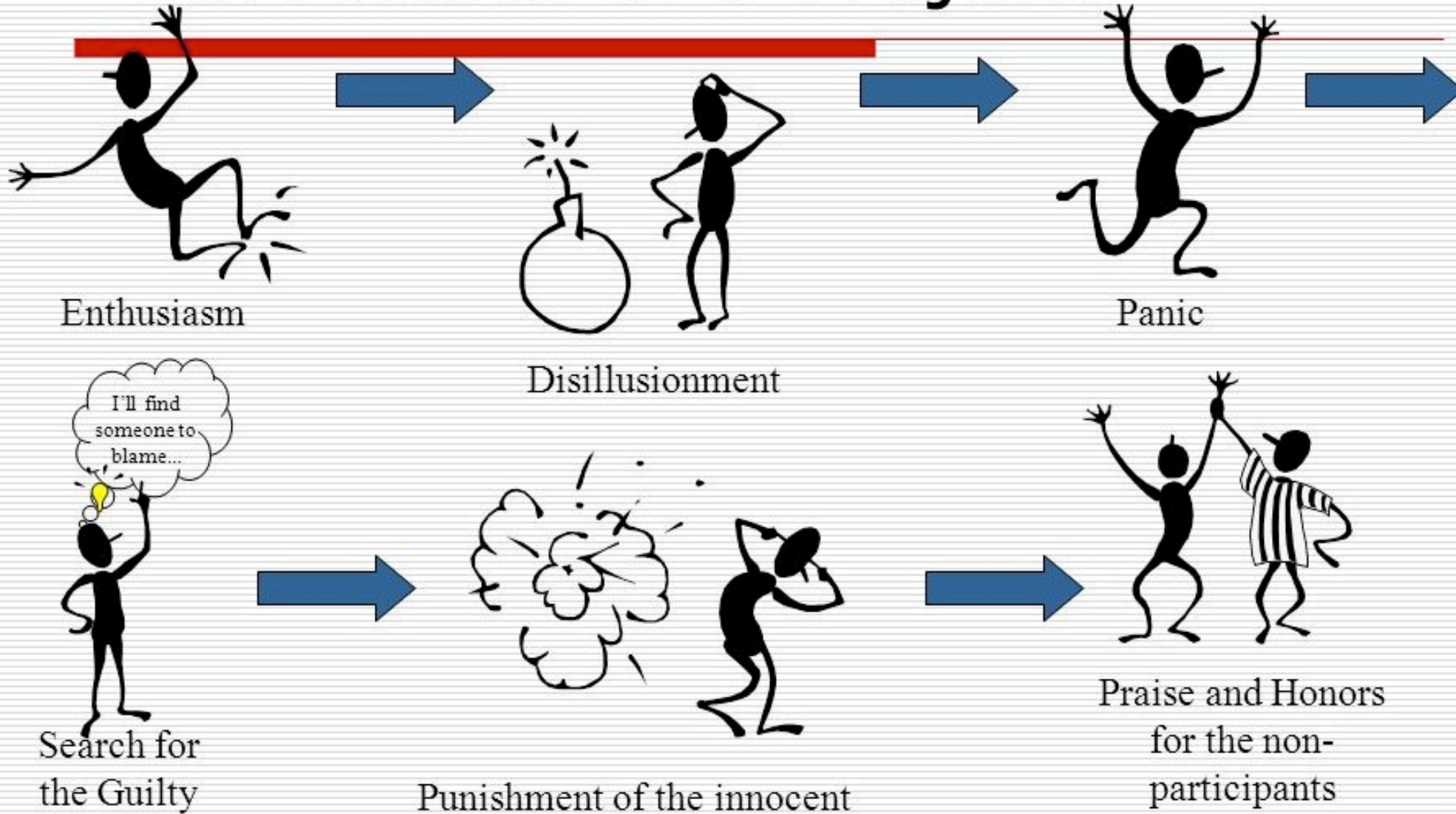


# What might the End Times be like?





# Six Phases of a Project



# More without Moore

- Real metric = better end user experience
  - Faster, cheaper, better, easier to use
  - Don't care how
- Picking the right problem is hard
- Not all problems are suited for academia (v8.1)
  - System – unlikely...
    - Far atomics, dirty bit, ...
  - Workloads – yes
    - SSRDMARHH (or whatever) is
- *Post talk discussion... V8.1*
  - *Yale points out that atomics and dirty bits were invented long time ago. ARM agrees. We're added optimizations for important use cases.*
  - *Atomics*
    - *Already had atomics, added far atomics.*
  - *Dirty bit*
    - *Already had a SW managed dirty bit, added HW managed dirty bit.*
  - *These types of optimizations will be difficult to research in academia as production SW and scaled out HW are required.*



# Potential Arch Research (Stuart's list)

1. Replace C with something better. The "replace" is important - needs to replace C's sizeable coder bum on seats advantage.
2. Keep cracking the parallel nut. The worlds programmers (outside of academia) are not very good at it. New languages? Use ML to extract it on programmers behalf?
3. More without Moore. General purpose compute will continue to be important. We want more performance without linear area/energy/power trade off. Not all can be addressed with heterogeneous or parallel compute. With demise of scaling do we revisit/rekindle some past ideas that unlock ILP?
4. Heterogeneous. Industry will be interested in accelerating applications it knows about. For the ones it doesn't frameworks are more valuable. E.g. Schemes to express the problem using popular languages, low friction routes to partition, schedule, resource share, virtualise, migrate work, optimise. How to generalise and make approaches relevant to more applications - are these techniques relevant to bare metal coders or apps developers?
5. Security. Keep progressing secure communication. Also progress exploit mitigation; replacing C would be a good start. We don't like C, did we mention that?
6. Reliability. Apply this to heterogeneous compute. Is there a framework or set of properties that participants need to adhere to?
7. New workloads, applications, new programming approaches and disciplines (e.g. approximate compute, stochastic compute). Look at these, tell us what's new and what won't work with existing ideas. Try not to just add a neural net to everything.



# Funnels

*Are it real and what about compute in memory?*

# Streaming Data. It's real and FUNDABLE...



## Geospatial Graph Analysis



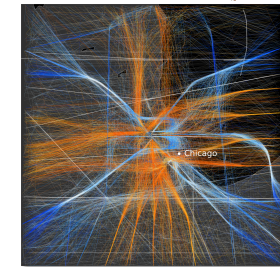
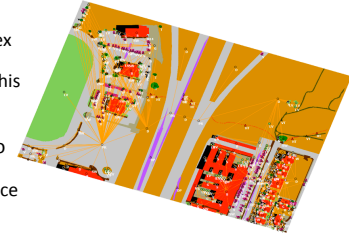
### Reality

- Sensors produce enormous quantities of complex data
- Current analysis capabilities fail to fully exploit this data to produce actionable intelligence

### Challenge: leverage the structure of geospatial data to identify patterns of life

- Automated data analysis capabilities that enhance human decision-making
- Scalable analysis over disparate temporal and geospatial scales
- Pattern analysis of complex trajectories

R&D is required at all levels of the software/hardware stack to automate the capture, fusion, and analytics of geospatial data streaming from heterogeneous sensors.



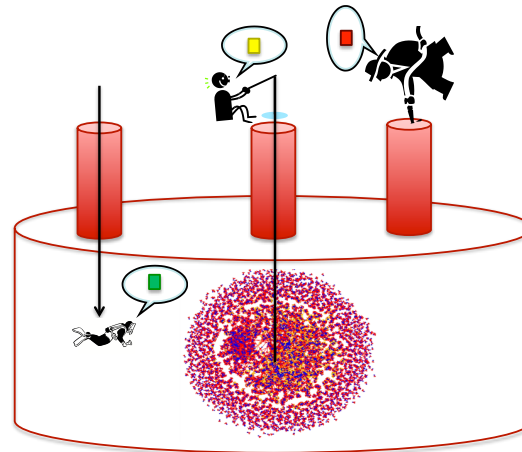
A convergence of HPC and graph-analytics is necessary to provide time-sensitive, actionable intelligence



2

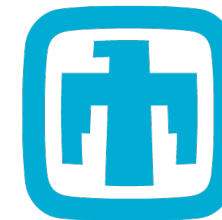
Approved for Unlimited Release: SAND2016-1943 O

## Big Data Graph Streaming



- Point query
- Fast ingest
- Global slice & dice, Data mining

■ Fault tolerance, code transparency, flexibility, data residence



**Sandia  
National  
Laboratories**



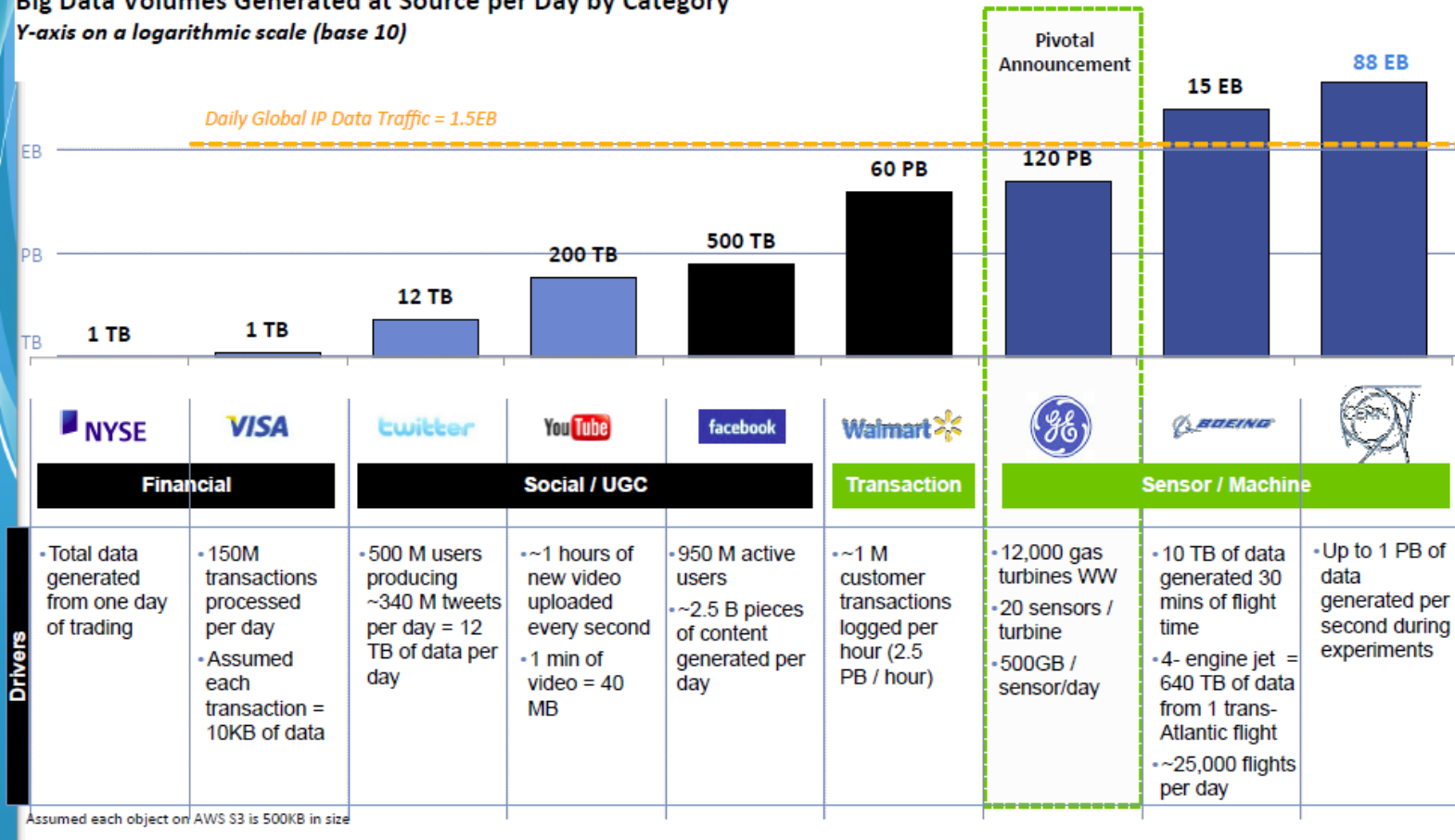
1

Approved for Unlimited Release: SAND2016-1943 O

# And the data will flood the Earth and kill us all...

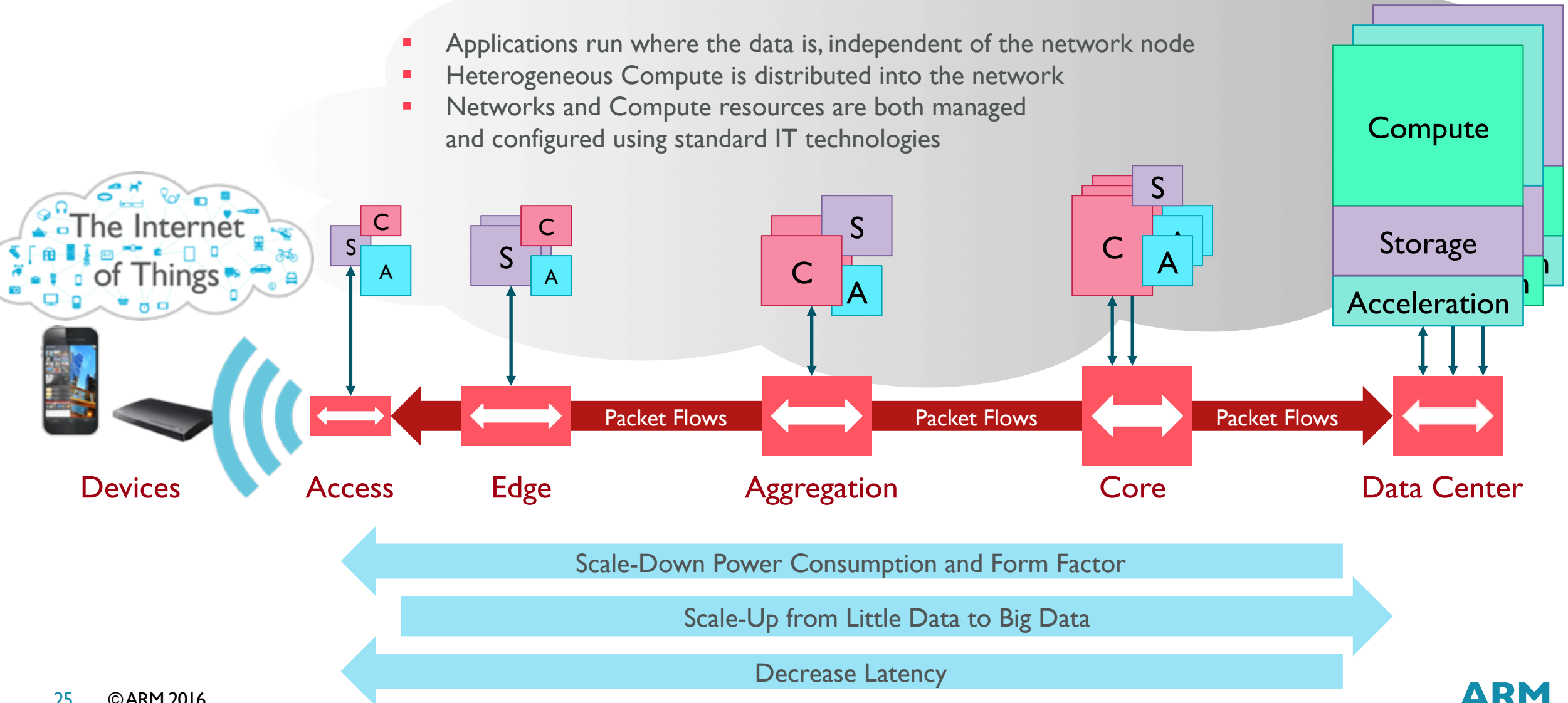
## IoT Big Data Generators Will Dwarf Platform Big Data Sources Like Video and Social!

Big Data Volumes Generated at Source per Day by Category  
Y-axis on a logarithmic scale (base 10)



# Still a funnel, Marketing's just better at Powerpoint

- Applications run where the data is, independent of the network node
- Heterogeneous Compute is distributed into the network
- Networks and Compute resources are both managed and configured using standard IT technologies



# Funnel I – On Chip



- On chip funnels
  - Also called embedded systems.
  - Embedded processor / microcontroller + RAM + Flash + accelerators + sensors + output
- Potential Research – start at the edge
  - Domain specify accelerators matched to sensor / data types
  - Detect interesting data
  - Compress data / extract metadata
  - Examples – can use FPGAs / embedded boards for platform
    - Keyword spotting (microcontroller/DSP)
    - Image processing (look inside the ISP) – ARM/Apical Spirit – video to metadata
  - Add security and/or safety for extra credit
- Impact of Research – ideas get absorbed into commodity platforms

# Funnel 2: Off Chip - Compute in Memory

## Power – Stay on-die!

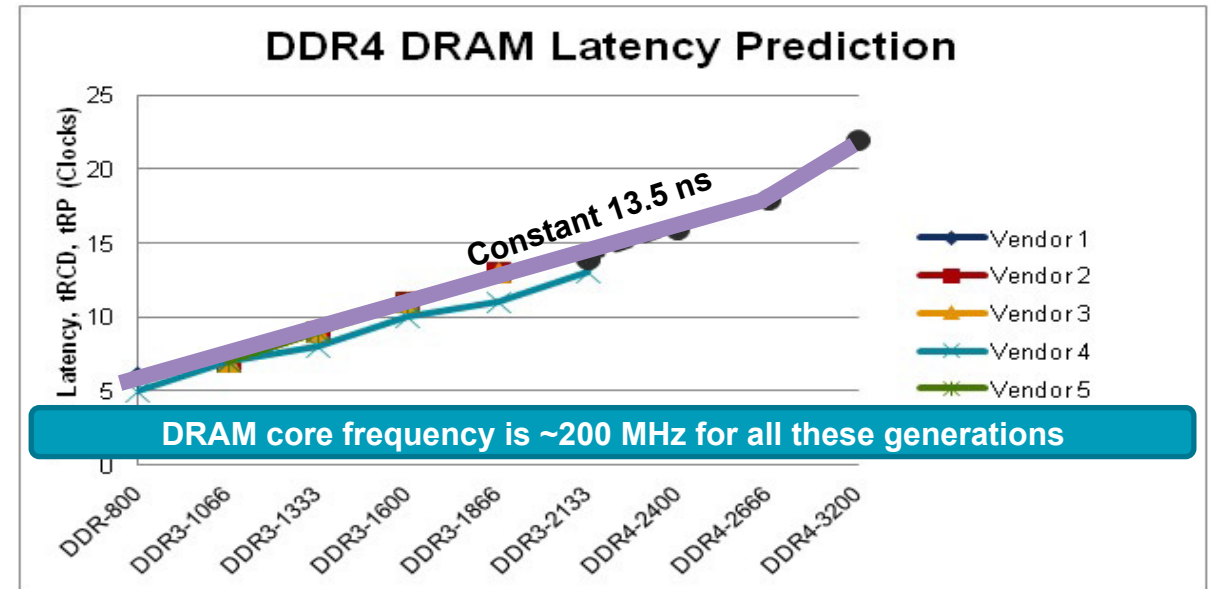
100 pJs gets you				
Cortex-M0	flash	On-chip mem	Off-chip mem	Bluetooth LE
10 cycles	Write 1 bit	Write 300 bits	Access 5 bits	Transmit .02 bits

## And a few other minor constraints

- Different processes for memory verse logic
- Memory is a commodity market

## Memory – Split across dies!

- Locality is split across multiple die BY DESIGN
- Memory is still 200Mhz, just more muxes



Source: Cadence (from DRAM manufacturers datasheets)

# Compute in Memory Continued

- You know Flash is dying... physics problems.
- Opportunities
  - NVM – likely need more compute anyway to hide how deep this stack of turtles really is.
  - Smarter memory controllers
  - Prefetching
  - Data reordering



# Potential Memory Research (Stephan's list)

1. tighter coupling with big.Little cores (we have one PhD student working on that), e.g., the cores sharing caches / branch predictors etc more closely, but not as crazy as some of the composite cores concepts
2. Non-volatile memory: how do the applications change for persistency? Hint: the existing proposals are not good enough. What does that mean for Endurance?
3. Future memory consistency models: what happens when we have different types of compute engines sharing the same memory? Are weak models the right thing? Can we do without coherence for ALL data, and instead have only coherent synchronisation variables?
4. Future memory architecture: we have had page-tables for ages; this day and age they are neither good for protection (for that we want to use things such as CHERI aka capabilities / regions) nor for remapping (see the pain with nested page tables, people trying to bring back flat mapped segments); and we have loads of indirection everywhere. If memory and storage merge, is Inodes / filenames everywhere the right thing? Why cannot everything that stores data live in a single address space? And be addressed via loads / stores E.g. network, register file, ..?
5. Future compute: hit the single-general-purpose performance wall; parallelise and specialise. What are “generic specialised” compute engines? E.g. GPU (we know those), but what else? Finite State Machines? Neural Network Accelerators? Data flow compute engines? How will they interact with the memory hierarchy? With the normal CPUs?
6. General: what are the next big ticket items? **Not 5% more**, because of doubling the complexity of the directory. Where are the step changes?
7. How do we fix the opposing parallelisation vs worsening QoS issue? In the limit, even *\*without\** a sequential bottleneck, the QoS of a parallel application will get worse, if the result can only be produced once all threads complete (due to the increased impact of jitter). Therefore, is there an aim to bring back wait-free compute? Approximate compute such that some parallel threads do not have to complete in time, but their result can be guessed (e.g. interpolate missing pixel values?)



## Commodity at scale\*

*big funnels or lots of funnels, that is the question*

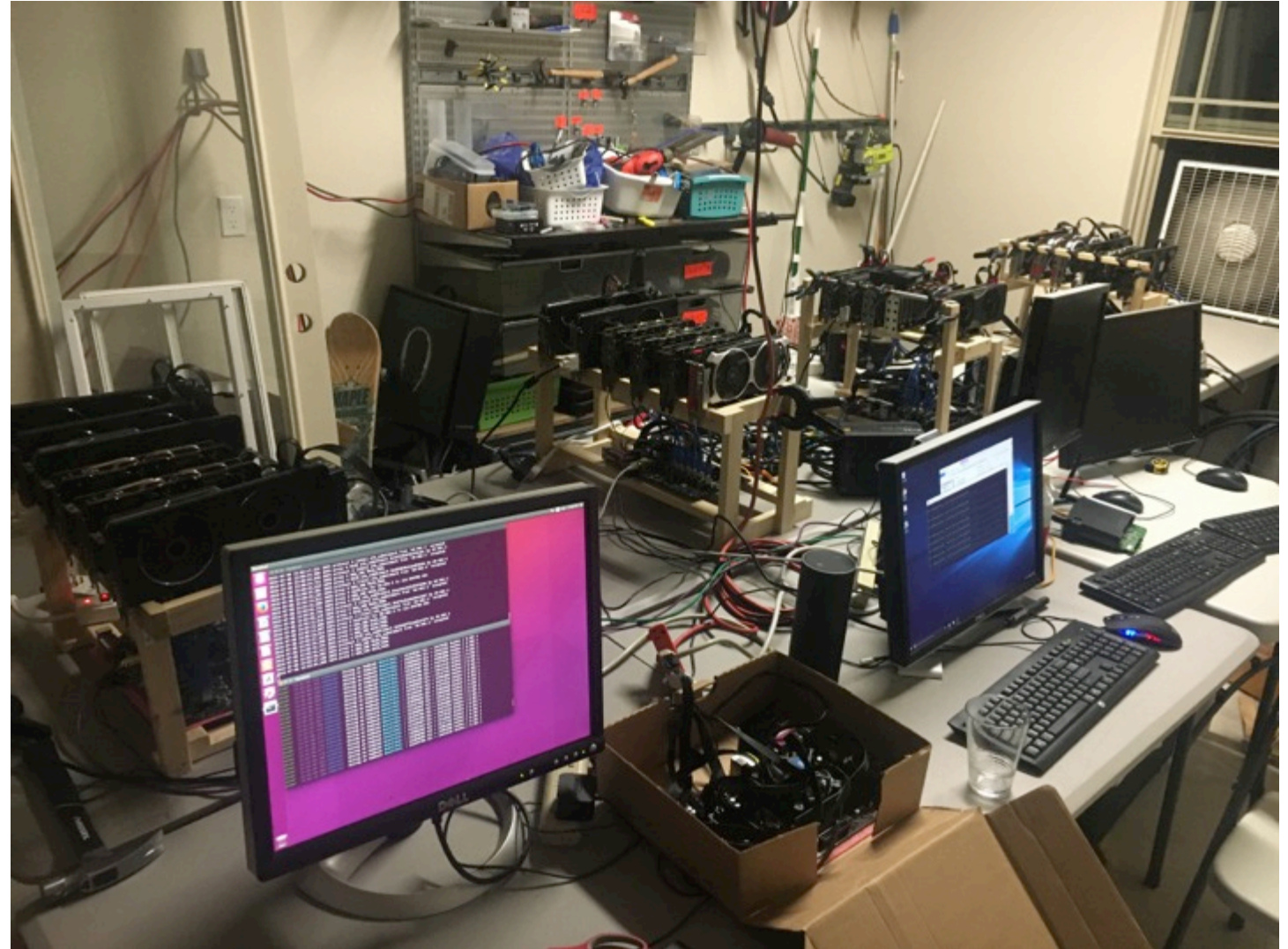
*How do you build mission critical compute infrastructure for a \$1 B startup using unreliable and untrusted compute? Assume \$0 funding for capital costs.*

### Problem Constraints

- Cern is 88 EB/day.... Let's start with that baseline
- Use only Amazon Prime & Ebay.
- Extra points if the system is buildable and maintainable by 12 year olds with basic YouTube skills

# Unreliable Computing Pilot (even better than approximate computing!)

- Compute nodes have
  - No redundancy
  - No climate control (42 C)
  - No ISO, no disaster recovery plan, no insurance, no contracts, no service level agreements....
- Node failures are common
  - Rain
  - Cats, kittens
  - Travel
  - Forced Windows updates
  - Random reboots



# Funnel 3: Commodity at scale

- Streaming data proxy app
  - Ethash (a modified version of Dagger-Hashimoto)
  - Memory hard / ASIC resistant
  - 5.7 MH/s = 50 GB/s per Tflop\*
- Ethereum PoW hashrate(July 18<sup>th</sup>)
  - 3,450 EB/day (CERN was 88EB/day)
  - 67,095 Eflops/Day

Network HashRate Growth Chart



# Ethereum PoW Compute: Cost break down

- 4,437 GH/s ETHash July 18<sup>th</sup>
  - 221,876 \* R9 380 GPUs (28nm) = Roughly \$61M all in, GPUs roughly 50%
  - Cost \$100K/day in power (50Mwatts)
  - 251 days ROI – July 18<sup>th</sup>
  - 41 days ROI – March (only 50K GPUs)
  
- Mining profitable in China until 1,000,000 GPUs in pool
  - 302K Efloper / day
  - \$1 / Efloper / day

Ethereum payout 5 ETH every 14 sec 1 ETH = \$11.7 (\$10-\$20 since Feb)	
Minute	\$250.71
Hour	\$15,042.86
Day	\$361,028.57
Week	\$2,527,200.00
Month	\$10,981,285.71
Year	\$131,775,428.57

- What else could be done if we use the HW we already have?
- Did I mention the \$50M Ethereum autonomous agent exploit. Google “theDAO”



# Post talk slide – commodity at scale

- *What's the point*
  1. *mainframe / big coherent computers (expensive) were disrupted by racks of message passing commodity servers.*
  2. *CERN @ 88EB / day looks scary, but is very doable today*
- *What's the opportunity?*
  - *Creating custom Si is hard and expensive, but there are other paths...*
  - *What if all those things we need to run data centers weren't really necessary (AC, ...)*
  - *How can we use commodity HW to create different types of systems with lower costs?*
  - *Ethereum (\$1B cryptocurrency) is clearing financial transactions by sharing compute from a network of untrusted and unreliable computers... all made possible by combining technology in new ways.*  
*Blockchain = p2p, Merkel trees, cryptographic hash,..*

**ARM**

University Enablement

# University Enablement (Free Stuff)

- Educational materials - <https://www.arm.com/support/university/>
- Research starter kits – launch in H2 2016
  - Gem5
  - Accelerators
- IP for research use
  - Tools - yes
  - Physical IP – short request form
  - Cortex-M0 – 149 universities. Free for evaluation and academic use
  - Other IP – available, contact ARM Research



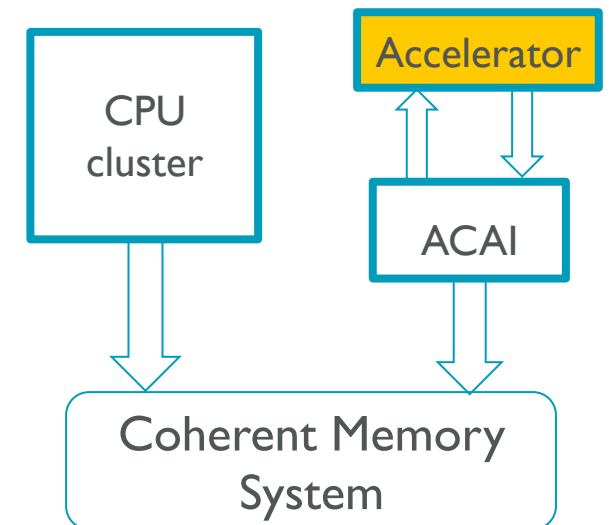
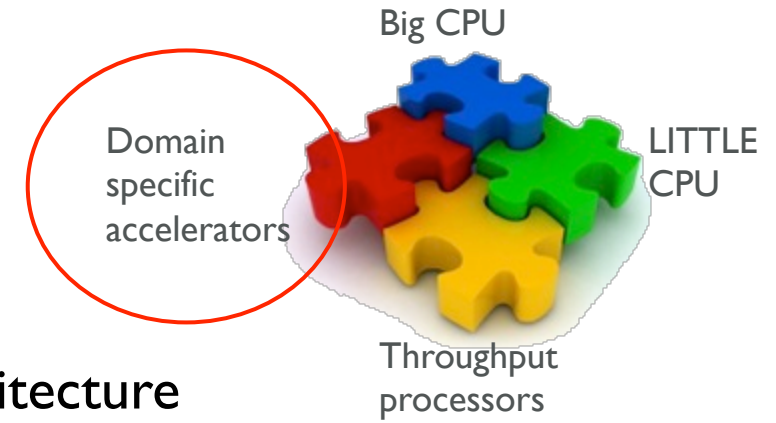
# Collaborations

- **ARM Research Summit – Sept 15<sup>th</sup>-16<sup>th</sup> Cambridge UK**
  - Event to boost the strategic collaboration links between universities and ARM
  - 10 invited speakers, 70 regular talks, 100+ attendees from Europe, US, Asia and South America
  - <https://developer.arm.com/research/summit>
- **Internship positions at ARM Research**
  - Several company summer internship positions in Cambridge/Austin offices (<http://www.arm.com/careers/students.html>)
  - Co-funded internships through the Hipeac programme in Cambridge office (<https://www.hipeac.net/mobility/internships>)
- **Participation in EC-funded Collaborative Projects**
  - Have participated in many EC-funded projects from FP6 to H2020 programmes (e.g. Hipeac 1-4, EuroCloud, MontBlanc 1-3)
  - Will participate in upcoming H2020 FETHPC & ICT calls in 2016/17
    - Contact: Emre Ozer - [emre.ozar@arm.com](mailto:emre.ozar@arm.com)



# Heterogeneous Compute

- Accelerators are a key component of heterogeneous compute
  - New instructions are nice but only so many ways to combine two source registers
- Established accelerators integrate with mainstream system architecture
  - c.f. HSA for GPU
- Standard accelerator integration requires cache maintenance, physical memory, device driver work
- ACAI: experimental hardware and software framework enabling easy adoption
  - Simpler programming model supporting user mode job dispatch, cache coherence & VA.
  - Developer writes user code
  - Standard drivers & shared libraries
- Chat with Stuart if you're interested





# Security and Specifications (Alastair's list)

## ARM Is

- Putting more security support in the hardware
  - Developed “TrustZone for ARMv8-M”
  - Security enhancements for IoT processors
- Putting reliability support in the hardware
  - Triple Core Lock Step ARM for Space project (Horizon 2020)
  - <http://www.tcls-arm-for-space.eu>
  - (ARM has provided commercial products with Dual Core Lock Step for 5 years)
- Making code and specifications publicly available
  - ARM's mbed uVisor publicly available
  - <https://github.com/ARMmbed/uvisor>
  - Public release of machine readable ARMv8 architecture specification this fall
  - Enable academic security work (and other systems software research)

# ARMv8 machine readable spec(Alastair's list)

## What's in the spec?

- Machine readable specification
- ARM, Thumb and 64-bit ISA
- System level specification (page tables, exceptions, ...)
- Hypervisor support
- TrustZone support
- System registers: fields, behaviour, access traps, banking
- Open source license
  - patent rights to build hardware explicitly not granted

## What could I do with it?

- Generate machine code analysis tools
- Build ARM simulator directly from the specification
- Build instrumented simulator to track information leaks
- Fuzz test bare metal software
- Augment software model checkers with model of hardware security features
- Check LLVM generates correct machine code
- Formally verify OS code: page tables, etc.
- Something we haven't even thought of

# Parting Thoughts for the Students

- There is no “good” work and “bad” work, there is just what you like to do
  - Our industry is full of brilliant and driven people.
  - You will only be successful if you’re doing what you love / passionate about.
  - Things you like to do that do not pay are called hobbies, keep looking
- Companies are like families and share a common culture
  - Multiple internships - play the field before you marry into one.
  - “INTEL’S COLLEGE HIRING METHODS AND RECENT RESULTS”, Bob Colwell...
    - RECOMMENDATIONS AND CONCLUSIONS
      - A – Intel’s design team interview process seems to work at the hire/don’t-hire level, but is a weak predictor of future success beyond that
      - B – Neither GPA, nor advanced degrees, nor school attended, are compelling predictors of success.
  - **ARM Research’s growth plan is to prioritize hiring successful interns. Looking for fit.**



**ARM** το τέλος