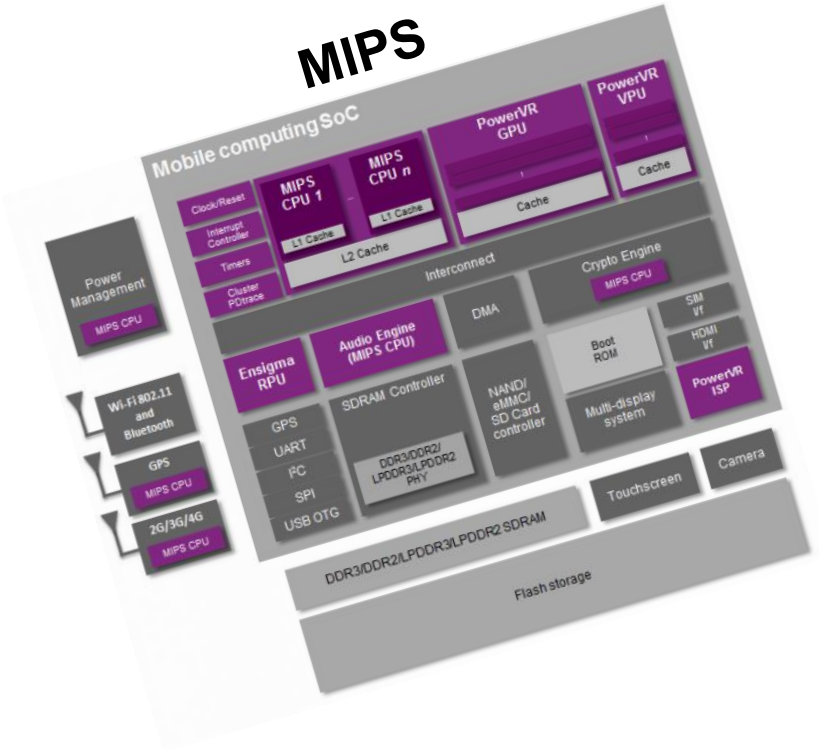
Part 2.1

Mobile Graphics Trends: Hardware Architectures

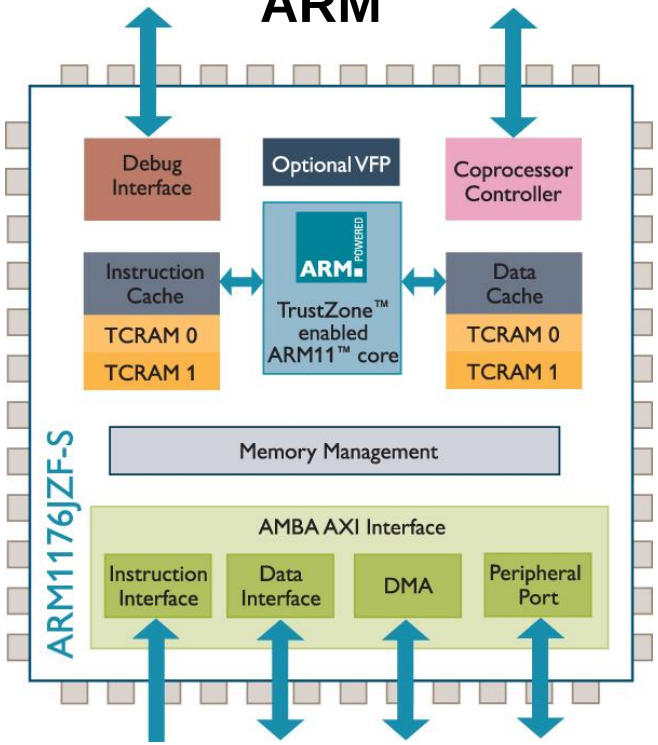
Pere-Pau Vázquez, UPC

Architectures (2014 – beginning 2015)

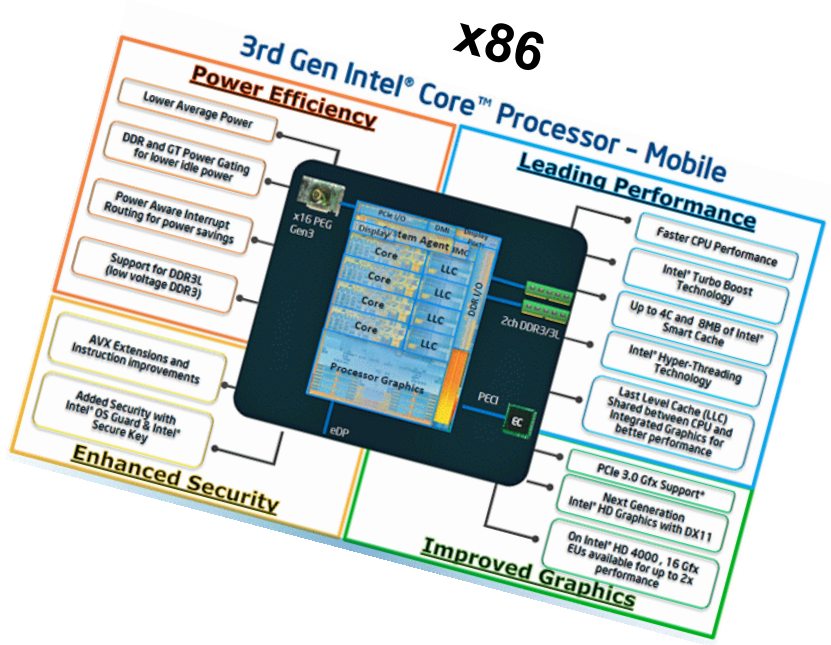
MIPS



ARM



x86



Architectures

- **x86 (CISC 32/64bit)**
  - Intel Atom Z3740/Z3770, X3/X5/X7
  - AMD Amur / Styx (announced)
  - Present in few smartphones, more common in tablets
  - Less efficient
- **ARM**
  - RISC 32/64bit
    - With SIMD add-ons
  - Most common chip for smartphones
  - More efficient & smaller area
- **MIPS**
  - RISC 32/64bit
  - Including some SIMD instructions
  - Acquired by Imagination, Inc. @2014

Architectures – RISC vs. CISC but…

- **CISC (Complex Instruction Set Computer)**
  - Fast program execution (optimized complex paths)
  - Complex instructions (i.e. memory-to-memory instructions)

- **RISC (Reduced Instruction Set Computer)**
  - Fast instructions (fixed cycles per instruction)
  - Simple instructions (fixed/reduced cost per instruction)

- **FISC (Fast Instruction Set Computer)**
  - Current RISC processors integrate many improvements from CISC: superscalar, branch prediction, SIMD, **out-of-order**
  - Philosophy → fixed/reduced cycle count/instr
  - Discussion (Post-RISC):
    - http://archive.arstechnica.com/cpu/4q99/risc-cisc/rvc-5.html

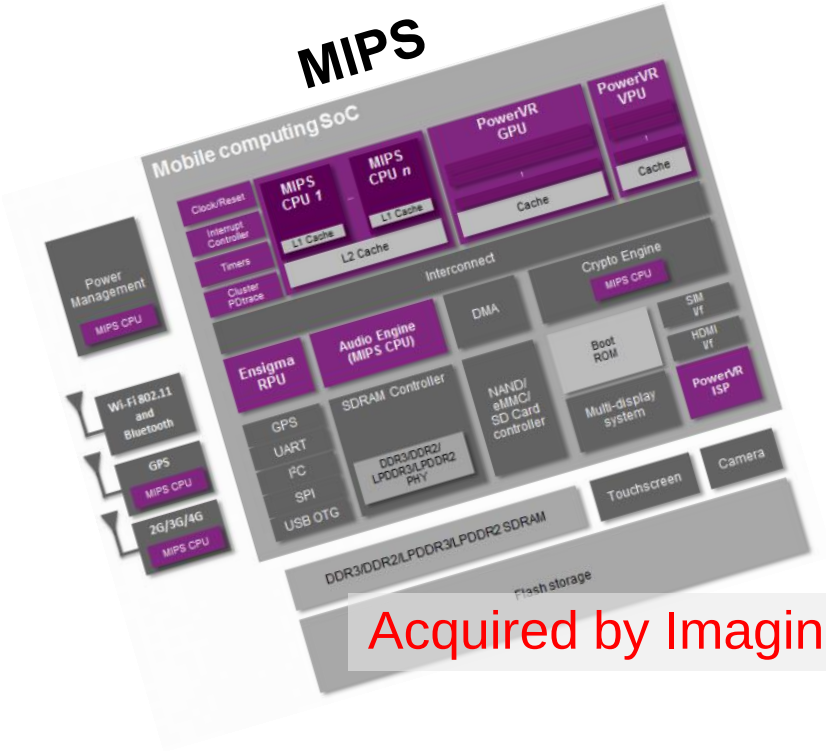
Landscape has changed a bit…

- **Status by 2014-2015:**
  - Intel Atom X3/X5/X7 announced (March 2015)
  - AMD announces Amur / Styx (20nm, Oct. 2014)
  - Nvidia launches Tegra X1 (March 2015)
  - ARM the only EU big technology company
  - Imagination announces Furian (sub 14nm, March 2017) Imagination's chips are in iPhones & iPads

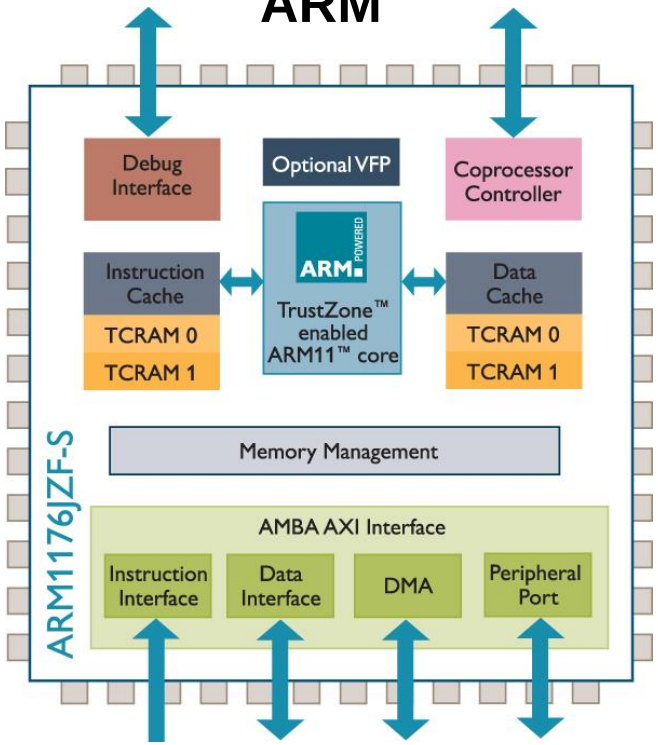
- **Nowadays:**
  - Intel quits mobile Apr/May 2016
  - AMD cancels 20nm chips (Jul. 2015)
  - Nvidia cancels Shield tablet (Aug. 2016)
  - ARM acquired by Softbank (Sep. 2016)
  - Apple tells Imagination that their IP will not be needed in 18-24 months (Apr. 2017)
    - Imagination sold to chinese-backed fund Canyon Bridge (Nov. 2017)
  - Broadcom has offered to pay $103 billion for Qualcomm
    - Declined
    - Broadcom is considering improving the offer (22th November)

Architectures (nowadays)

MIPS



ARM



Acquired by Imagination, inc

x86

Architectures – ARM

- **ARM Ltd.**
  - RISC processor (32/64 bit)
  - IP (intellectual property) – Instruction Set / ref. implementation
  - CPU / GPU (Mali)
- **Licenses (instruction set OR ref. design)**
  - **Instruction Set** license -> custom made design (SnapDragon, Samsung in Galaxys, Apple in iPones & iPads)
    - Optimizations (particular paths, improved core freq. control,…)
  - **Reference design** (Cortex A9, Cortex A15, Cortex A53/A57…)
- **Licensees (instruction set OR ref. design)**
  - Apple, Qualcomm, Samsung, Nvidia, AMD, MediaTek, Amazon (through Annapurna Labs, Inc.)…
  - Few IS licenses, mostly adopting reference design
- **Manufacturers**
  - Contracted by Licensees
    - GlobalFoundries, United Microelectronics, TSM…

Architectures – ARM…

- **Suppported on**
  - Android, iOS, Win Phone, Tizen, Firefox OS, BlackBerry, Ubuntu Phone, …
- **Biggest mobile market share**
- **Typically paired with mobile GPUs. Existing brands:**
  - Adreno 4x0/5x0 – Qualcomm
  - PowerVR 8XE (Rogue) – Imagination
  - Mali T8x0/G51/G71 – ARM
- **General strategies:**
  - Cache coherence – week sequential code guarantees on multithreading!!
  - Heavy **dependence on compiler** → optimize instruction scheduling
    - Operation dependencies , loop unrolling, etc…
  - Use SIMD extensions

Architecture types

- **High performance**
  - Premium smartphones & tablets
- **High area efficiency**
  - Medium-to-low smartphones
- **Ultra-low power**
  - Smartwatches

Architectures

Mobile GPU architecture trends

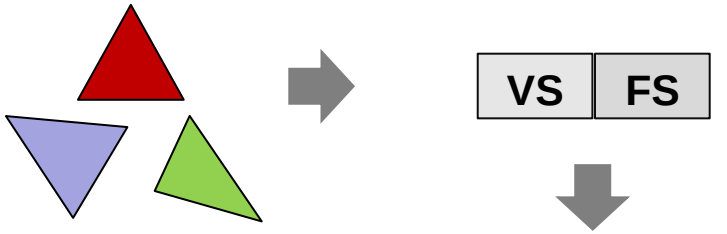
Graphics pipeline trends

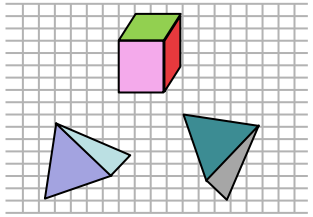
- **Tiled rendering**
- **Data (texture) compression**
- **Other optimizations**

Tiled Rendering

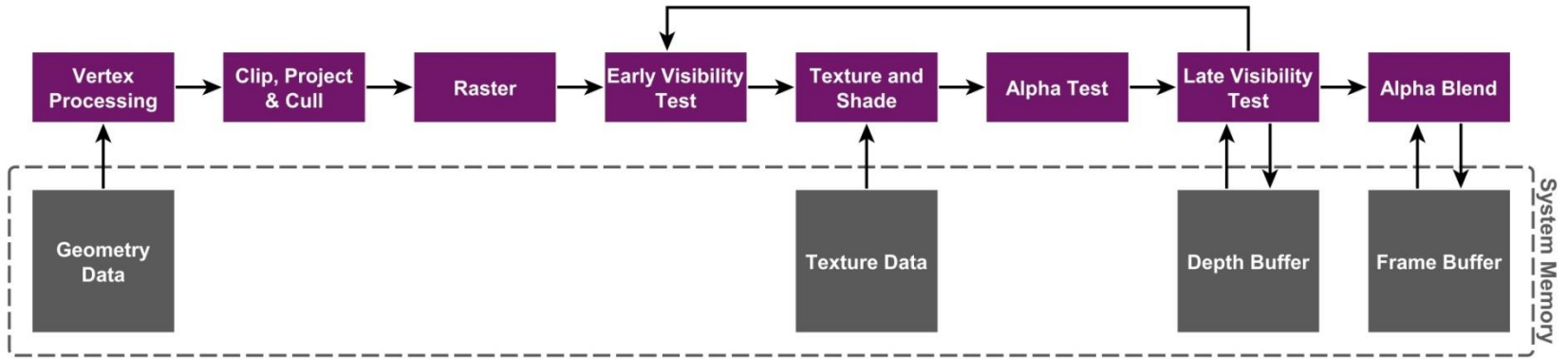
- **Immediate Mode Rendering (IMR)**
- **Tile-Based Rendering (TBR)**
- **Tile-Based Deferred Rendering (TBDR)**

Architectures – GPU



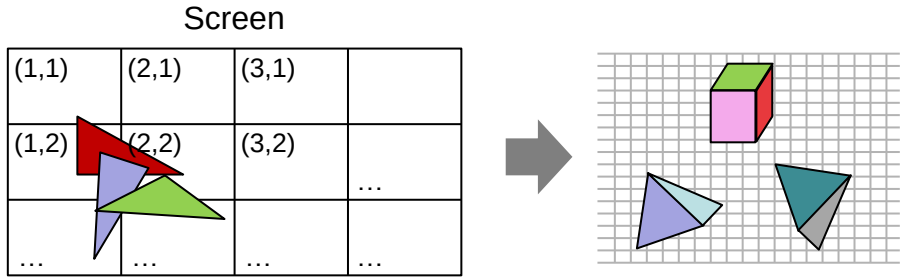


- **Inmediate Mode Rendering (IMR)**
  - Geometry is processed in submission order
    - High **overdraw** (shaded pixels can be overwritten)
  - Buffers are kept in System Memory
    - High bandwidth / power / latency
  - Early-Z helps depending on geometry sorting
    - Depth buffer value closer than fragment → discard

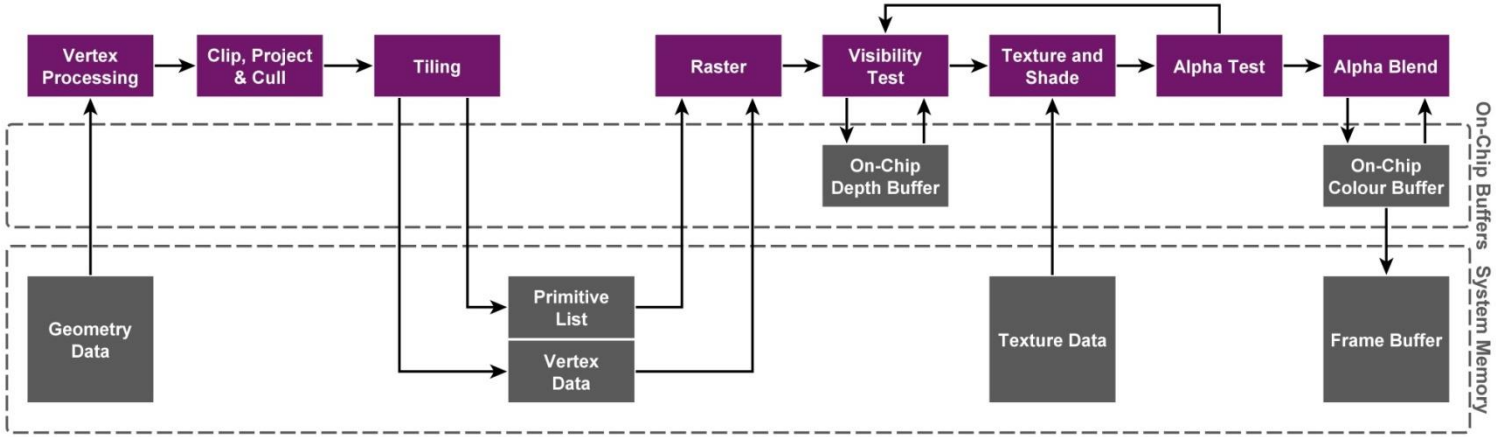


http://blog.imgtec.com/powervr/understanding-powervr-series5xt-powervr-tbdr-and-architecture-efficiency-part-4

Architectures – GPU

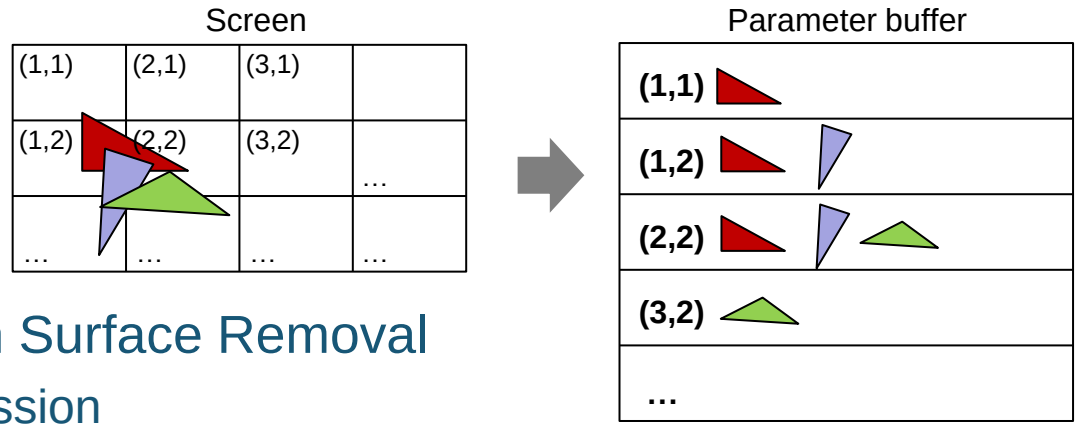


- **Tile Based Rendering (TBR)**
  - Rasterizing per-tile (triangles in bins per tile) 16x16, 32x32
    - Buffers are kept on-chip memory (GPU) – fast! → **geometry limit?**
  - Triangles processed in submission order (TB-IMR)
    - **Overdraw (front-to-back -> early z cull)**
  - Early-Z helps depending on geometry sorting

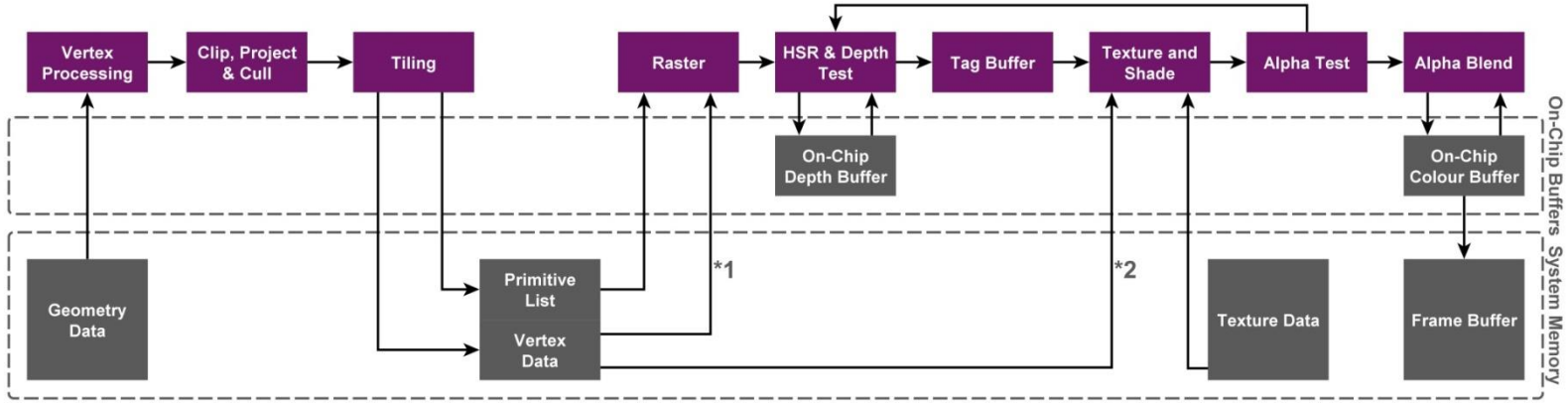


http://blog.imgtec.com/powervr/understanding-powervr-series5xt-powervr-tbdr-and-architecture-efficiency-part-4

Architectures – GPU



- **Tile Based Deferred Rendering (TBDR)**
  - Fragment processing (tex + shade) ~waits for Hidden Surface Removal
    - Micro Depth Buffer – depth test before fragment submission
      - whole tile → **1 frag/pixel** ☺ ⟶ **Limit: ~100Ktri + complex shader**
    - iPAD 2X slower than Desktop GeForce at HSR (FastMobileShaders_siggraph2011)
  - Possible to prefetch textures before shading/texturing
  - Hard to profile!!! ~~~Timing?



http://blog.imgtec.com/powervr/understanding-powervr-series5xt-powervr-tbdr-and-architecture-efficiency-part-4

Data/texture compression

- ARM's Adaptive Scalable Texture Compression (ASTC) supported by most mobile GPU vendors
- ETC2/EAC standard compression OpenGL ES 3.0
- Compression hardware also present in display hardware
  - Rendered images stored and transferred to the display in a compressed
    - Saving bandwidth

Other optimizations

- Deferred shading
- Primitive elimination
- Skipping updates to pixels that do not change
  - ARM memory transaction elimination

Trends

- Specific hardware for ray tracing
- Deep learning libraries & hardware (e.g. Qualcomm's Fast CV, Nvidia's CUDA Deep Neural Network)
- Skipping updates to pixels that do not change
  - ARM memory transaction elimination