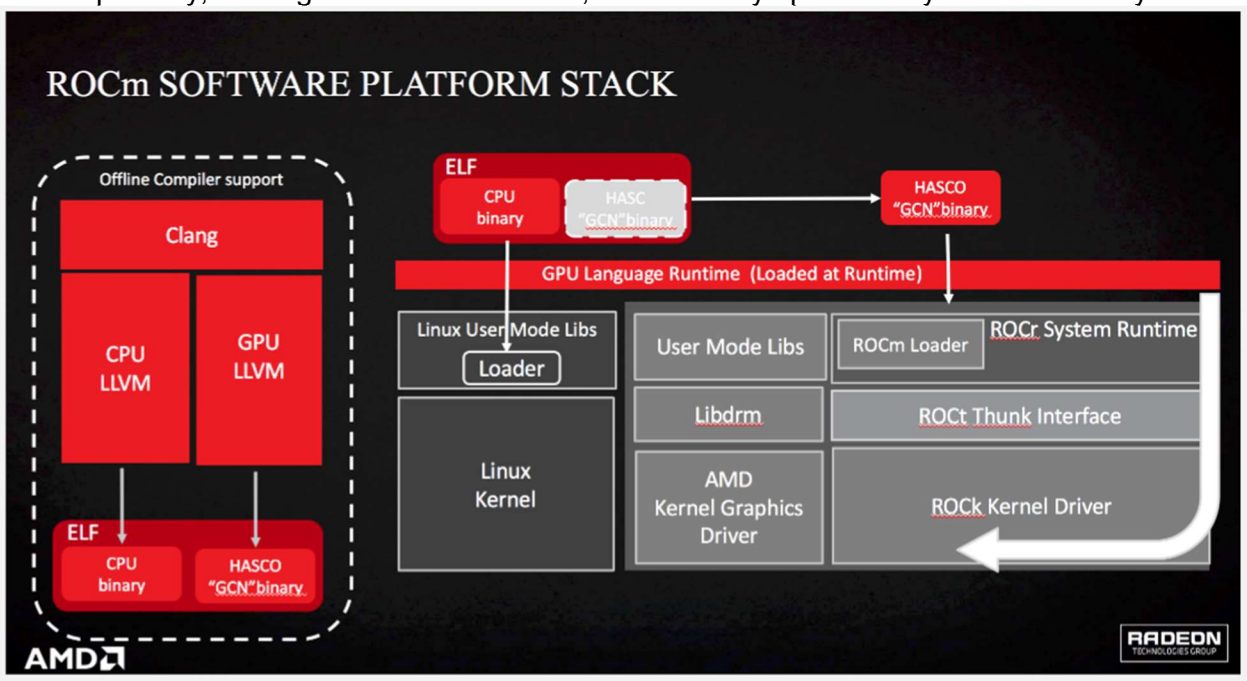
Notes related to ROCm

ROCM

Is a compiler, driver, and etc. that is aware of the equipment contained in a machine and optimizes the running of your code for you

A goal is to code something in one language Ex. C++ AMP, and have it execute portably, at a high level of abstraction, and in a very optimized system aware way



ISA: Instruction Set Architecture

The design of instruction set, ISA determines the semantic characteristics of the instruction set of a computer.

SA: System architecture

A system architecture is the conceptual model that defines the structure, behavior, and more views of a system.

HSA-Heterogeneous system architecture

Describes system conceptually, is considered an “initiative”

Make CPU’s and GPU’s work together seamlessly

High bandwidth access to memory, high performance, low power

Support for diverse set of high level programming languages

Applications will flow between processor types

Easier to convert existing programs

openCL, C++ AMP, OpenMP, Java, Python, Javascript

HSA becomes the framework upon which these languages can be built upon

“Democratizes compute”

ROCm is the realization of HSA in a multimachine transparent system

HSAIL is HSA intermediate language, assembly language of HSA

Is the lower level software implementation of HSA

Is mapped to the specific hardwares ISA that encapsulates many of the HSA

Shields the software stack from the specific hardware, delivers performance via the finalizer. This allows for write once run anywhere programmability

https://www.youtube.com/watch?v=Me0g_FsgU0U

Anything coded with OpenCL 2.0 or higher will be able to be run on HSA

Coding will theoretically be easier, as less focus is necessary on memory

LLVM
- Is a collection of compiler and toolchain technologies
- Contains OpenMP
- Consider LLVM the name for an alternative open source compiler set to GNU

Clang is a front end compiler which runs on top of the llvm (see compiler section below)
- LLVM is used as an optimizer and back end
- Compiling with “clang file.c” is effectively a new compiler in competition with gcc
- Clang is newer than gcc, and is constantly updated and supported
- It allows for some fancy compilation debugging, as gcc is more of a monolithic static compiler
- Clang supposedly has better error messages
- Clang runs faster and uses less memory than gcc
- It generates system independent LLVM IR code
- http://clang.llvm.org/comparison.html

HC: Heterogeneous Compute
- The concept of writing a highly abstracted single piece of C++ code which will automatically be appropriated to the GPU and CPU
  - There are many optimizations which will occur in the background
- HC is a C++ API which can be compiled with HCC

HCC: Heterogeneous Compute Compiler
- Compiler implementing the HC concept
  - Generates code for both CPU and the accelerator
- Must include <hc.hpp> header file
- Support two dialects
  - C++ AMP, HC, HIP
- Is based on and uses parts of LLVM/Clang
- Generates HSAIL or GCN ISA
- http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2015/p0069r0.pdf

HIP: Heterogeneous-Computing Interface for Portability
- HIP is a C++ dialect designed to ease conversion of CUDA applications into portable GPU code
- Hipify will convert CUDA to HIP code
- HIP can be run on AMD and NVIDIA devices

ROCR: Runtime
- Runtime manipulation is when you specify a specific run time action to be done
  - Ex. spawn a thread, CUDA is a runtime API
- ROCR is an API that allows you to do:
  - Error handling
  - Runtime initialization and shutdown
  - System and agent information
  - Signals and synchronization
  - Architected dispatch
  - Memory management

ROC-smi: system management interface
- Clock and temperature management

ROCK: Kernel
- Big picture: the kernel patch/driver that allows ROCm the ability to do what it does

Thunk:
- Seems to be some sort of communicator for the GPU driver

When machine is ready to do compiling you should see an executable for hcc-hsail and hcc-lc
- hcc-hsail will produce HSA IL code

This is being depreciated and should not be used
hcc-lc uses the GCN ISA backend and produces code for graphics core next processors
Supports Hawaii and Fiji family
GCN-Graphics core next
GPU architecture, with DirectX 11.2 support, trueaudio support, and mantle api
With an APU described as 4 cpu and 8 gpu units, the 8 is referring to GCN
Each GCN has 4 simd units, which each process 16 work items and are meant to take on a wavefront
A wavefront can be thought of as a warp in CUDA
Rocm requires a newer HSA and most likely won't work on older desktops/systems
hUMA-Heterogeneous uniform memory access
hQ-Heterogeneous queueing, how the scheduler works
Gets rid of traditional master slave relationship
https://www.youtube.com/watch?v=i6BWzL12KMI
https://www.youtube.com/watch?v=4YV6z6Fgw48
Github for ROCm:
https://github.com/RadeonOpenCompute
AMD posts about ROCm:
http://gpuopen.com/rocm-do-you-speaka-my-language/

Our Systems
Vulcan:
Has four AMD FX 8800P Radeon R7 12 Compute Core 4C+8G APUs
(APU - has CPUs and GPUs on the same die)
Stats stored in proc/cpuinfo
Libraries for rocm are stored in opt/rocm
ssh (first initial last name)@vulcan.ece.neu.edu

Cyclops:
Used to replace vulcan due to reliability issues
Also contains 4 APUs
ssh (first initial last name)@cyclops.ece.neu.edu
Still does not support ROCm because it is only APU, which ROCm was not designed for

ROCm
Has 64 AMD Opteron(tm) Processor 6366 HE (CPUs according to proc/cpuinfo)
Has 2 R9 Fury nano's and one ASPEED "VGA compatible controller" according to
sudo lshw -C display
ssh (first initial last name)@129.10.47.183

Other General Computer Notes
Building Programs:

This is a great intro guide to compiling and downloading dependencies
http://www.howtogeek.com/105413/how-to-compile-and-install-from-source-on-ubuntu/

CMake
Is a makefile generator
Run by "cmake (Path to CMakeLists.txt)"

You often have to set variables with flags (-D sets variables)

cmake -D Foo=1 MyProject/CMakeLists.txt # --> Works

cmake MyProject/CMakeLists.txt -D Foo=1 # --> Fails

https://www.youtube.com/watch?v=T4BiC24Y16Y

Autotools

If you ever see autogen.sh file

Run the following

$ sudo chmod +x autogen.sh

$ ./autogen.sh

$ ./configure --prefix=/some_directory

$ make

$ make install

What is apt-get?

apt=advanced packaging tool

Used to install new packages, remove, upgrade etc.

Designed to avoid dependency hell (will auto install deps)

Apt searches its list of cached packages located in ets/apt/sources.list

There are four major package repository types in Ubuntu:

main - Supported by Canonical. This is the major part of the distribution.

restricted - Software not licensed under the GPL (or similar software license), but supported by Canonical.

universe - Software licensed under the GPL (or similar license) and supported by users.

multiverse - Software not licensed under the GPL (or similar license), but supported by users.

There are also these additional types of repositories:

Trusty-updates - Updates to official packages.

Trusty-backports - Current version software from Terrific Tiger (Trusty+1) that have been backported to Trusty Tahr.

Trusty-proposed - Proposed updates & changes (bleeding edge stuff).

Third party repositories are not supported or verified by GPL or Canonical and may contain viruses

PPA is a Personal Package Archive and is basically just someone who is hosting it on their own, can destroy a computer so be careful

Options for sudo apt-get

sudo apt-get install

sudo apt-get install -f (fix an install)

sudo apt-get remove

Sudo apt-get download (just downloads .deb package)

sudo apt-cache search

Search for repo in listed repos in ets/apt/sources.list

sudo apt-get update

Update cache after adding/removing repos

sudo apt-get upgrade

Upgrade package

There are often many package options when doing a chache search. -dev packages contain the header files as well, necessary if you are linking manually to the library (developer)
Underneath it will download a .dep file and upac it in the directories specified by a contained config file
If you have a .deb file:

sudo dpkg -i packagename.deb (installs like apt-get install would)

This will not handle dependencies

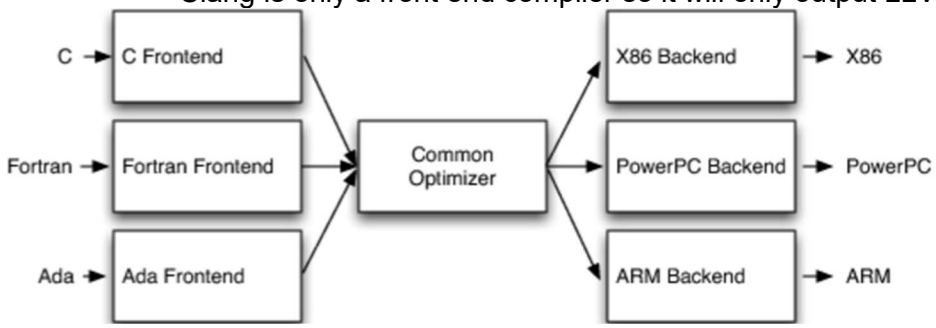
apt-config dump shows the system apt-get system set up
apt pinning allows for the forcing of a specific version to be downloaded
http://ubuntuguide.org/wiki/Ubuntu_Trusty_Packages_and_Repositories

Description of dpkg flags (-x is for extraction)
http://askubuntu.com/questions/40779/how-do-i-install-a-deb-file-via-the-command-line

Compilers

Compilers consist of 3 parts
Front end: lexical analysis, parsing
Optimizer: optimizing the code
Back end: machine code generation
In gcc these steps are all together so this concept is somewhat foreign

Clang is only a front end compiler so it will only output LLVM IR



Bios

Bios is a small chip/set of firmware that checks to make sure that all components are present and operational. Then control is handed off to the operating system

Encryption

You have your own public key which is sent to a server
The server then encrypts based on this key
You receive the message and decrypt based on your private key
Hashing is when you apply an algorithm and store a password as a complicated string, cannot be unlocked with a key
https://www.youtube.com/watch?v=dut9EnbFym0

Port Forwarding (also known as tunneling)

Connect to a specific computer by specifying the port number on a router in a LAN that the computer you are trying to communicate with is connected to

This conceptually tells you the location of the computer by going through the router

VPN
- Acts as if your computer is within the LAN
- Better security, all encrypted

32 vs 64 bit
- Refers to the RAM usable by the operating system
- 2^32 vs 2^64 (4 gig vs 16 billion gigs)

Types of connectors:
- VGA is an analog visual connection, suffers from signal degradation
  - 2 mp bandwidth
- DVI more bandwidth but no audio or data
- HDMI everything
- Displayport everything but even better and newer

Driver: Acts as an abstraction layer or translator
- New driver necessary for new piece of hardware and a new OS

DRAM ex. DDR3 is dynamic random access memory
- Cut power to it and stored data is lost
- Requires refreshing to keep memory stored within it
- Isn't used for long term storage

SRAM, cache
- Working memory for task at hand

Flash memory, ex. SDD
- Doesn't disappear when power is shut off, same use intention as a disk drive

GPUs
- OpenCL is used for AMD gpus
- Cuda is used for NVIDIA
- C++AMP: C++ Accelerated Massive Parallelism
  - New C++ wrapper supposedly allowing for very easy GPU computing

gpuOPEN
- New ways to access the GPU
- Started by AMD
- Libraries are often black boxes, no ability to manipulate source code
- Allows access to GPU libraries, open source
  - Most content is stored on Github
- Closer access to the GPU
- Interaction with developers, regular blog posts on AMD website
- https://www.youtube.com/watch?v=Y0oBFeFUG4w&list=PLx15eYqzJifefHhMHV9qRqsHhmbagRQG0

OpenPower
- Allow more design over the hardware components of the system being designing to fit needs

Power is an open source ISA and description of hardware that the companies are also hoping will lead to an advancement in the field

The Open Source aspect of the power movement applies to a specific subset of products

Ex. Power8 is the newest OpenPower CPU

P100 is also part of the series

IP: Internet Protocol

Standard communication system/set up

IP address is a hierarchical organization of numbers corresponding to country / regional network / sub-network / address of device

Transitioning from IPv4 to IPv6 --> need more IP addresses

IPv4 (Internet protocol version 4) is a 32 bit address (2^32 possible addresses)

IPv6 (version 6) is a 128 bit address (2^128 possible addresses) - developed in 1999, standardized in 2012

Information is communicated via IP packets

These take routes from the internet page server to your laptop in an often indirect route determined by internet traffic and quickness etc. (handled by routers)

MAC addresses are used to identify intermediate computers/servers in the path that the data is taking

Packets may arrive in unordered sequences

TCP (transmission control protocol) does an inventory on your computer's end sending acknowledgements of each piece of info to the sending server

Key concepts: many paths from server to host IP = reliability

HIP: Host Identity Protocol

Host identification technology

Internet comprised of two namespaces: IP addresses and Domain Name System (DNS)

DNS handles requests from computers (URL = uniform resource locator)

There are servers for each type of domain (.org .com etc.)

https://www.youtube.com/watch?v=5o8CwafCxnU

Google has its own DNS (8.8.8.8 or 8.8.4.4)

HIP separates end-point identifier and locator roles of IP addresses

Dynamic IP vs Static IP

IP's are assigned by the ISP (internet service provider)

Can be either static or dynamic

Dynamic IP: IP may change based on the needs of the service provider. This is the default setting - needed because there are more connected devices than available IPs.

Configured through the DHCP (dynamic host configuration protocol)

Static IP: IP remains fixed - necessary for addresses which must remain constant (like a server)

Linux:

root directory: the "/" directory, level closest to the OS of the system

access via "cd /"

home directory: each user has a home directory,

accessed by just calling the "cd" command.

the home directory is denoted by the "~" character, so paths can be written relative to the home directory.

Framework:

skeleton structure which needs to be filled in with "programmatic flesh" to make a complete application.
"Flesh" uses different pieces of the skeleton to make ends meet.
generally ship with a set of predefined functions
QT libraries are an example of a framework

Software:

digital instructions that guide the operation of computer hardware. general term referring to any of the code that runs on the computer.

Firmware:

a special type of software that is not intended to change once a product is shipped.
updates to firmware require either changing the chip or a special code that reloads the flash memory containing the software.

ELF: a file format for executables and shared object libraries

Translation unit: a c/cpp file after it has included all header files

SCP vs SFTP

SCP:

Only transfers files
Non interactive
Usually faster than SFTP
scp [options] username1@source_host:directory1/filename1 username2@destination_host:directory2/filename2

SFTP:

More interactive like create directories, deleting files etc.

Toolchain

Set of tools that are used for software development
Goal is to create an appropriate executable
    GNU is a toolchain
Toolchains are often used for an external device that may not be able to host the development environment (called a cross toolchain)
    Ex. Developing and testing for an embedded system
May contain a debugger or compiler for a specific language
Concepts: Host system, target system, emulate the target system on the host