Lecture 16
Class Project Introduction
and
Case Study: DianNao (Part 1)

Xuan ‘Silvia’ Zhang

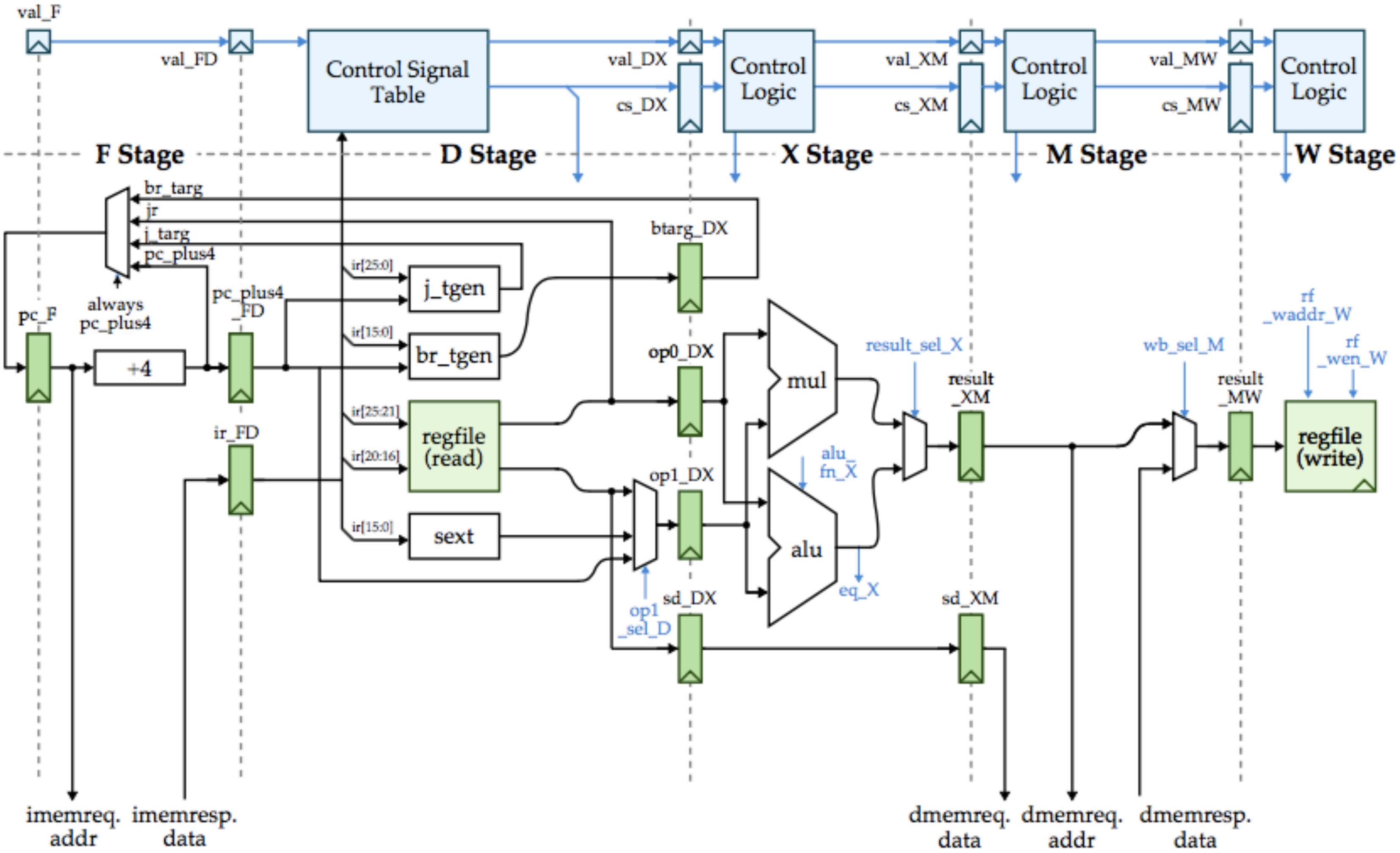
Washington University in St. Louis

http://classes.engineering.wustl.edu/ese566/

Quiz: Adding a New Auto-Incrementing Load Instruction

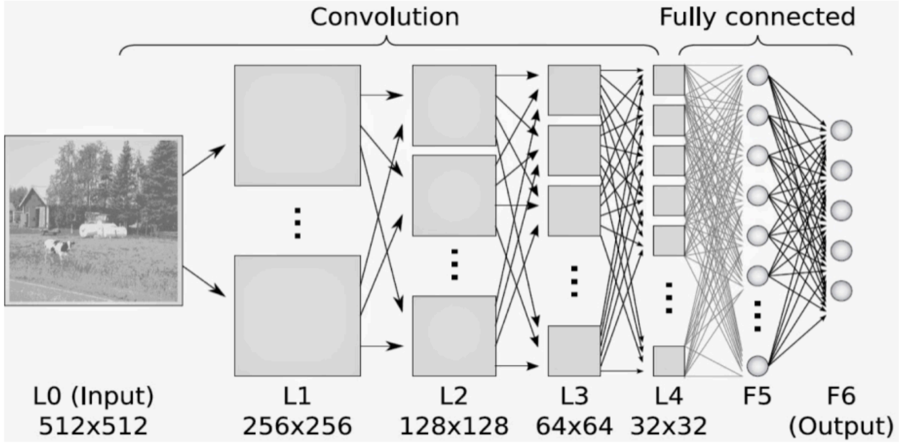
`lw.ai rt, imm(rs)`

$$R[rt] \leftarrow M[\,R[rs] + sext(imm)\,];\ R[rs] \leftarrow R[rs] + 4$$



Class Project Introduction

- Convolutional neural network (CNN)
  - an advanced artificial neural network algorithm
  - highly successful in image recognition applications



- Design a CNN hardware accelerator
  - latency and throughput
  - power and area

Project Timeline

- 3/22: Brief introduction; start forming teams
- 3/27: Release description; team finalized
  - http://classes.engineering.wustl.edu/ese566/Lab/ClassProject.pdf
- Week 11-12: Review related research papers
  - DianNao
  - Eyeriss
- 4/10: Submit initial project proposal/plan with block diagrams and interfaces
- Week 13-14: No lecture. Project team meetings
- 4/24 and 4/26: Project presentation
- 5/8: Final project report

How to “Research”

- Solve a non-trivial problem
  - somewhat unique
  - can’t immediately google the answer
  - require design thinking

- Research Process
  - open-ended
  - non-linear
  - iterative
  - creative

Research Process

- Orientation
  - what have been done? (google scholar)
  - where is the community? (conference and journals)
  - who are most active? (university research groups)
- Distillation
  - breakthrough ideas (survey/review paper)
  - seminal publications (highly cited paper)
  - fundamental theoretical concepts (textbook)
- Replication
  - detailed implementation (tools + datasets + methods)
  - inquire original authors
- Innovation
  - address bottleneck
  - take on high-impact challenges

How to Read Research Papers

- Browse Broadly (<5min per paper)
  - google scholar alert
  - follow conference proceedings and journal issues
  - glance at title, maybe abstract, sometimes conclusion
- Read Selectively (~30min per paper)
  - intro: motivation
  - figures and results: competitive performance
  - techniques and methods: how to replicate?
  - evaluations: is it fair?
- Review Intensively (> weeks per paper)
  - understand everything
  - trace all important reference
  - learn the story telling and the logic organization

DianNao (ASPLOS 2014)

- ASPLOS
  - ACM International Conference on Architectural Support for Programming Languages and Operating Systems
  - other examples: ISCA, MICRO, HPCA, ISSCC, JSSC, DAC
- Ubiquitous Machine Learning
  - artificial neural network (ANN)
  - convolutional neural network (CNN)
  - deep neural network (DNN)
- High-Throughput
  - large-scale neural network
  - impact of memory

Reading Assignment Questions

- Team 1
  - Andrew Ellison, Shixuan Zhang
- Team 2
  - Brett Gilpin, Matthew Wedewer, Nestor Gonzalez
- Team 3
  - Weidong Cao, Xinyao Li, Liu Ke
- Team 4
  - An Zou, Meizhi Wang, Longzhen Zhang
- Missing Slides
  - Bo, Ding, Hu, Huang, Li, Li, Liu, Yin

Project Proposal

- Due on 4/10 at noon

- Conceptual implementation
- Block diagrams of the system
- Well-defined interfaces and design parameters
- Target performance specification
- Division of work among team members

Questions?

Comments?

Discussion?