Introduction to Embedded Systems Research
Final Exam

Robert Dick

28 April 2020

Open notes, papers, and internet. With the exception of asking clarifying questions of the instructor, do not communicate with others about the exam.
Make sure the photo or scan you send is legible.
There are answer length constraints for some questions; those are strict.

Due at 3:30pm 28 April.

Name:

Sign below to acknowledge the Engineering Honor Code: "I have neither given nor received aid on this examination, nor have I concealed a violation of the Honor Code."

$\boxed{4}$ 1. The executives at a startup company have access to marketing consultants. These consultants are more experienced with the customer discovery interview process than the executives, and the executives are uncomfortable talking with potential customers. What is the most important disadvantage (or disadvantages) of hiring the marketing consultants to conduct the customer discovery interviews, instead of having the executives conduct the interviews themselves? Use at most three sentences.

$\boxed{3}$ 2. In Jerry Liu's work on fingertip tracking, what was the primary mechanism considered to trade off computational efficiency and accuracy?

- ○ Varying neural network depth.
- ○ Varying neural network activation function.
- ○ Spatial subsampling.
- ○ Temporal subsampling.
- ○ Changing the amount of available memory.

$\boxed{1}$ 3. Which parameter of a computer system is most directly modified during voltage scaling? Use at most three words. Don't overthink this.

$\boxed{3}$ 4. What prevents designers from changing this parameter to a value resulting in near-zero energy consumption? Use at most three sentences.

$\boxed{2}$ 5. In recent years, the proportion of energy consumed for global communication on an integrated circuit has increased relative to that for computation. In what specific part of an integrated circuit is most of energy for communication used? Use at most three words.

$\boxed{2}$ 6. An embedded system uses an iterative sub-sampling process akin to Digital Foveation. It has two cameras, one of which is capable of sub-sampling, and the other of which is only capable of full-resolution sampling. Both cameras have maximum resolutions of 1600x1200 RGB pixels with eight bits per pixel color channel. Capture requires $T_c \times n$ time, where n is the number of bits transferred. Data transfer from image sensor to applications processor requries $T_t \times n$ time. Analysis requires $T_a \times n$ time. The capture, transfer, analysis process must occur 60 times per second.

- $T_c = 20\,\mathrm{ps/bit}$,

- $T_t = 80\,\mathrm{ps/bit}$, and
- $T_a = 200\,\mathrm{ps/bit}$.

What is the utilization of the applications processor if the full-resolution camera is used? Do not assume pipelining: the processor is occupied and therefore not available for other activities during capture and transfer so the total time will be the sum of capture, transfer, and analysis times. Two significant digits are sufficient.

2 7. What is the utilization of the applications processor if the subsampling camera is used, capturing and operating on 10% of the data in the full-resolution case in each iteration, but requiring two iterations per inference event? Again, don't assume pipelining so the total processor time is the sum of capture, transfer, and analysis times. Two significant digits are sufficient.

4 8. If the applications processor has four cores and analysis has perfect parallelization efficiency, having one faulty core would result in a 1.01 utilization for the applications processor given maximum-resolution non-foveated capture and analysis, indicating that system timing constraints cannot be met in this case. Faulty cores only increase analysis time, not the other times. For all combinations of functioning

Table 1: System-Level Fault State

Functional CPU cores	Subsampling camera	Utilization (fill these in)
4	functional	you answered this above
4	faulty	you answered this above
3	functional	
3	faulty	1.01
2	functional	
2	faulty	
1	functional	
1	faulty	

components indicated in Table 1, indicate the utilization.

3 9. Y. Chen, N. Chiotellis, L.-X. Chuo, C. Pfeiffer, Y. Shi, R. G. Dreslinski, A. Grbic, T. Mudge, D. D. Wentzloff, D. Blaauw, and H. S. Kim, “Energy-autonomous wireless communication for millimeter-scale Internet-of-Things sensor nodes,” *IEEE J. on Selected Areas in Communications*, vol. 34, no. 12, Dec. 2016.

Why is the energy for transmit pulses drawn from the capacitor instead of the battery? Use at most two sentences.

4 10. Y. Zhu, A. Samajdar, M. Mattina, and P. Whatmough, “Euphrates: Algorithm-SoC co-design for low-power mobile continuous vision,” arXiv, Tech. Rep., Apr. 2018.

Describe the purpose of the Motion Controller, and indicate why its tasks aren’t, instead, executed on the CPU? Use at most three sentences.

4 11. The frequencies of biological neuron spiking events communicate information. A neuron’s behavior can be modeled, coarsely, as integrating input spikes, and generating an output spike when the integral exceeds a threshold. Indicate a more complex mechanism existing in biological spiking neural systems that is not captured by the basic input integration model described above. Use at most two sentences.

5 12. P. Coussy, C. Chavet, H. Wouafo, and L. Conde-Canecia, "Fully binary neural network model and optimized hardware architectures for associative memories," *ACM J. on Emerging Technologies in Computing Systems*, vol. 11, no. 4, Apr. 2015.
Write an expression for the number of bits of weight memory required for a network as a function of the number of clusters, C, and fanals, L. State any assumption that are not clear in the paper.

4 13. E. Ronen, A. Shamir, A.-O. Weingarten, and C. O'Flynn, "IoT goes nuclear: Creating a ZigBee chain reaction," in *Proc. Symp. on Security and Privacy*, May 2017.
What were the most significant security flaws exploited by the authors? You may consider both hardware and software design and implementation. Use at most four sentences.

$\boxed{4}$ 14. S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "EIE: Efficient inference engine on compressed deep neural network," in *Proc. Int. Symp. Computer Architecture*, June 2016. Han et al. used a varient of compressed sparse columns to encode sparce weight matrices. Why do they insert a zero in vector v when the number of zeros in a run exceeds 15? Are there design alternatives? What are their advantages and disadvantages of their approach? Use at most four sentences.

$\boxed{5}$ 15. Describe the concept of Hyperdimensional Computing. This concept was not covered in the course, but several methods to quickly find and understand new research concepts were covered in detail. Use at most three sentences.

Thank you all for taking the course so seriously and sticking with it despite the difficult circumstances. You did better than I could ask or expect and brightened an otherwise dark semester.