Theme 2 Demonstrator – Runtime Optimisation of Stereo Matching

INTRODUCTION

- Demonstration of the adaptive runtime management of a dynamic application on an embedded platform: Stereo Matching on the Odroid-XU3.
- Model-based prediction and control of knobs (threads and frequency) and optimisation of monitors (power and perf.) through the PRiME Framework.
- Application- and platform-agnostic runtime management for power and performance optimisation.
- Optimisation of power whilst meeting a performance requirement.

APPLICATION LAYER

- Stereo Matching is depth estimation
- Found in dynamic environments e.g. autonomous driving
- Processing from ZED Stereo Camera
- Dynamic application frame-rate based on the depth of objects in the scene
- Multi-threaded in pthreads and OpenCL
- 2 algorithms for compute vs accuracy trade-off

- Algorithm {GIF, SGBM}
- Algorithm mode (SGBM only) {0,1,2}
- Max disparity \([1, \infty)\)
- Resolution {376,720,1080,1242}
- Compute mode {pthreads (CPU), OpenCL (GPU)}
- Number of threads \([1, \infty)\)
- +7 tuning parameters for SGBM

APPLICATION MONITORS

- Performance (FPS)
- Number of pixel errors
- Mean disparity/depth error

RUNTIME MANAGEMENT LAYER

- Demonstration of the adaptive runtime management of a dynamic application on an embedded platform: Stereo Matching on the Odroid-XU3.
- Model-based prediction and control of knobs (threads and frequency) and optimisation of monitors (power and perf.) through the PRiME Framework.
- Application- and platform-agnostic runtime management for power and performance optimisation.
- Optimisation of power whilst meeting a performance requirement.

- Odroid-XU3: Quad-core A7 & A15 clusters + GPU (OpenCL 1.2 on GPU)
- Power & temperature sensors on SoC

- Requirements of applications and configuration of platforms cannot be predicted at design time.
- Create a runtime management approach that is sufficiently accurate, adaptable and has a low training overhead.

RESEARCH CHALLENGES

- Multiple Linear Regression Modelling
- Predict monitor value from given set of knob values
- Online model-building using a minimum training samples for error convergence

- Runtime optimisation - gradient descent
- Search problem - use models to find knob settings that meet performance requirement and minimise power
- Repeat when app. requirements change

DEV ICE LAYER

- Odroid-XU3: Quad-core A7 & A15 clusters + GPU (OpenCL 1.2 on GPU)
- Power & temperature sensors on SoC

- Frequency
- Power
- CPU core affinity
- Temperature

Charles Leech: c.leech@soton.ac.uk

Application Knobs

- Algorithm {GIF, SGBM}
- Algorithm mode (SGBM only) {0,1,2}
- Max disparity \([1, \infty)\)
- Resolution {376,720,1080,1242}
- Compute mode {pthreads (CPU), OpenCL (GPU)}
- Number of threads \([1, \infty)\)
- +7 tuning parameters for SGBM

Application Monitors

- Performance (FPS)
- Number of pixel errors
- Mean disparity/depth error

Application layer

App knobs

App monitors

App-to-RTM API

Runtime Management layer

RTM-to-Dev API

Dev knobs

Dev monitors

Device layer

Power and Performance Scaling

PS: 32 cores 952 MHz
P6: 30 cores 667 MHz

Charles Leech: c.leech@soton.ac.uk

Application Knobs

- Algorithm {GIF, SGBM}
- Algorithm mode (SGBM only) {0,1,2}
- Max disparity \([1, \infty)\)
- Resolution {376,720,1080,1242}
- Compute mode {pthreads (CPU), OpenCL (GPU)}
- Number of threads \([1, \infty)\)
- +7 tuning parameters for SGBM

Application Monitors

- Performance (FPS)
- Number of pixel errors
- Mean disparity/depth error

Application layer

App knobs

App monitors

App-to-RTM API

Runtime Management layer

RTM-to-Dev API

Dev knobs

Dev monitors

Device layer

Power and Performance Scaling

PS: 32 cores 952 MHz
P6: 30 cores 667 MHz

Charles Leech: c.leech@soton.ac.uk

Application Knobs

- Algorithm {GIF, SGBM}
- Algorithm mode (SGBM only) {0,1,2}
- Max disparity \([1, \infty)\)
- Resolution {376,720,1080,1242}
- Compute mode {pthreads (CPU), OpenCL (GPU)}
- Number of threads \([1, \infty)\)
- +7 tuning parameters for SGBM

Application Monitors

- Performance (FPS)
- Number of pixel errors
- Mean disparity/depth error

Application layer

App knobs

App monitors

App-to-RTM API

Runtime Management layer

RTM-to-Dev API

Dev knobs

Dev monitors

Device layer

Power and Performance Scaling

PS: 32 cores 952 MHz
P6: 30 cores 667 MHz

Charles Leech: c.leech@soton.ac.uk

Application Knobs

- Algorithm {GIF, SGBM}
- Algorithm mode (SGBM only) {0,1,2}
- Max disparity \([1, \infty)\)
- Resolution {376,720,1080,1242}
- Compute mode {pthreads (CPU), OpenCL (GPU)}
- Number of threads \([1, \infty)\)
- +7 tuning parameters for SGBM

Application Monitors

- Performance (FPS)
- Number of pixel errors
- Mean disparity/depth error

Application layer

App knobs

App monitors

App-to-RTM API

Runtime Management layer

RTM-to-Dev API

Dev knobs

Dev monitors

Device layer

Power and Performance Scaling

PS: 32 cores 952 MHz
P6: 30 cores 667 MHz

Charles Leech: c.leech@soton.ac.uk

Application Knobs

- Algorithm {GIF, SGBM}
- Algorithm mode (SGBM only) {0,1,2}
- Max disparity \([1, \infty)\)
- Resolution {376,720,1080,1242}
- Compute mode {pthreads (CPU), OpenCL (GPU)}
- Number of threads \([1, \infty)\)
- +7 tuning parameters for SGBM

Application Monitors

- Performance (FPS)
- Number of pixel errors
- Mean disparity/depth error

Application layer

App knobs

App monitors

App-to-RTM API

Runtime Management layer

RTM-to-Dev API

Dev knobs

Dev monitors

Device layer

Power and Performance Scaling

PS: 32 cores 952 MHz
P6: 30 cores 667 MHz

Charles Leech: c.leech@soton.ac.uk

Application Knobs

- Algorithm {GIF, SGBM}
- Algorithm mode (SGBM only) {0,1,2}
- Max disparity \([1, \infty)\)
- Resolution {376,720,1080,1242}
- Compute mode {pthreads (CPU), OpenCL (GPU)}
- Number of threads \([1, \infty)\)
- +7 tuning parameters for SGBM

Application Monitors

- Performance (FPS)
- Number of pixel errors
- Mean disparity/depth error

Application layer

App knobs

App monitors

App-to-RTM API

Runtime Management layer

RTM-to-Dev API

Dev knobs

Dev monitors

Device layer

Power and Performance Scaling

PS: 32 cores 952 MHz
P6: 30 cores 667 MHz

Charles Leech: c.leech@soton.ac.uk

Application Knobs

- Algorithm {GIF, SGBM}
- Algorithm mode (SGBM only) {0,1,2}
- Max disparity \([1, \infty)\)
- Resolution {376,720,1080,1242}
- Compute mode {pthreads (CPU), OpenCL (GPU)}
- Number of threads \([1, \infty)\)
- +7 tuning parameters for SGBM

Application Monitors

- Performance (FPS)
- Number of pixel errors
- Mean disparity/depth error

Application layer

App knobs

App monitors

App-to-RTM API

Runtime Management layer

RTM-to-Dev API

Dev knobs

Dev monitors

Device layer

Power and Performance Scaling

PS: 32 cores 952 MHz
P6: 30 cores 667 MHz

Charles Leech: c.leech@soton.ac.uk