ASPEN: Acceleration of Visual-Inertial Odometry for Extended Reality on an FPGA

Kathleen Feng

Stanford University

31 August 2023
Extended Reality

**Augmented Reality (AR):** overlay of digital information on real world, supplements reality

**Virtual Reality (VR):** full immersion into computer-generated environment, opaque displays and sensory input

**Mixed Reality (MR):** manipulation of both physical and virtual items, interaction with real world and virtual environment

**Extended Reality (XR):** umbrella term for AR+VR+MR
Extended Reality Pipeline

Perception Pipeline
- Hand/Eye Tracking
- Visual-Inertial Odometry
- Scene Reconstruction

Visual Pipeline
- Application Rendering
- Post-Processing
- Display

Audio Pipeline
- Recording
- Encoding
- Playback

Camera, IMUs, Depth Camera
Mic
Constraints for an Ideal System

Three axes to evaluate XR system:

1. Performance:
   - Motion-to-photon (MTP) latency: under 20 ms (VR), 7 ms (AR)
   - Throughput: ideally 100 gigapixels/s

2. Power:
   - Power budget: under 1-2 W (VR), 0.1-0.2 W (AR)

3. Quality:
   - Accuracy in subtasks: visual-inertial odometry, eye gaze estimation
   - User experience: interaction, responsiveness,...
Constraints for an Ideal System

Three axes to evaluate XR system:

1. **Performance:**
   - Motion-to-photon (MTP) latency: under 20 ms (VR), 7 ms (AR)
   - Throughput: ideally 100 gigapixels/s
Constraints for an Ideal System

Three axes to evaluate XR system:

1. Performance:
   - Motion-to-photon (MTP) latency: under 20 ms (VR), 7 ms (AR)
   - Throughput: ideally 100 gigapixels/s

2. Power:
   - Power budget: under 1-2 W (VR), 0.1-0.2 W (AR)
Constraints for an Ideal System

Three axes to evaluate XR system:

1. Performance:
   • Motion-to-photon (MTP) latency: under 20 ms (VR), 7 ms (AR)
   • Throughput: ideally 100 gigapixels/s

2. Power:
   • Power budget: under 1-2 W (VR), 0.1-0.2 W (AR)

3. Quality:
   • Accuracy in subtasks: visual-inertial odometry, eye gaze estimation
   • User experience: interaction, responsiveness,...
## Current State of Commercial XR Systems

<table>
<thead>
<tr>
<th>Feature</th>
<th>Ideal</th>
<th>Varjo VR-3</th>
<th>Quest 2</th>
<th>HoloLens 2</th>
<th>Quest Pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage</td>
<td>–</td>
<td>VR</td>
<td>VR</td>
<td>MR</td>
<td>MR</td>
</tr>
<tr>
<td>Resolution (MPixels)</td>
<td>400</td>
<td>15.7</td>
<td>7.0</td>
<td>4.4</td>
<td>6.9</td>
</tr>
<tr>
<td>Refresh Rate (Hz)</td>
<td>240</td>
<td>90</td>
<td>90</td>
<td>120</td>
<td>90</td>
</tr>
<tr>
<td>MTP latency (ms)</td>
<td>&lt; 20 (VR)</td>
<td>&lt; 20</td>
<td>N/A</td>
<td>&lt; 9</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>&lt; 7 (AR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Power (W)</td>
<td>&lt; 1-2 (VR)</td>
<td>N/A</td>
<td>N/A</td>
<td>&gt; 7</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>&lt; 0.1-0.2 (AR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mass (g)</td>
<td>100-200</td>
<td>944</td>
<td>503</td>
<td>566</td>
<td>722</td>
</tr>
</tbody>
</table>
CPU Processing Time Breakdown

Source: ILLIXR [Huzaifa 2021]
VIO: Visual-Inertial Odometry

Calculates 3D user position from sensors

- IMUs, cameras $\Rightarrow (x, y, z, \theta, \phi, \psi)$
- Most dominating subtask, represents $\sim 40\%$ of XR workload

Using OpenVINS as gold model [https://docs.openvins.com/index.html]
VIO: Visual-Inertial Odometry

- Visual Pipeline
- IMU Pipeline
- State Update

- Left image
- Right image
- IMU measurements

- camera rate (20-30 Hz)
- IMU rate (200 Hz)
Visual Pipeline

1. Extract features (if needed) from previous frame
2. Track features between previous and current frame
Visual Pipeline

1. Extract features (if needed) from previous frame
2. Track features between previous and current frame
IMU Pipeline

Provides fast state updates between camera frames

[Diagram description]

- IMU Updates
- Calculate noise covariance, state transition matrix
- Sum state ($\Phi$, $Q$)
- EKF propagation (new state)
- Augment state with new clone

Previous State → State
State Update

Features two update paths for two sets of features
1. MSCKF features: newer, freshly extracted features
2. SLAM features: features that have appeared in several consecutive frames
State Update

Features two update paths for two sets of features

1. MSCKF features: newer, freshly extracted features
2. SLAM features: features that have appeared in several consecutive frames

```
(2)
```
FPGA Acceleration

Currently prototyping accelerator on Xilinx Versal VCK190 FPGA

• Using Vitis HLS to implement vision and IMU pipelines on reconfigurable fabric
• On-board application processor performing all other computation
• In the process of testing end-to-end application on FPGA
Visual Pipeline Optimizations

- Switched almost everything to integer
- Absolute trajectory error: difference between the estimated trajectory and groundtruth after it has been aligned

$$e_{ATE} = \sqrt{\frac{1}{K} \sum_{k=1}^{K} ||x_{k,i} - \hat{x}^+_{k,i}||^2}$$

<table>
<thead>
<tr>
<th></th>
<th>Orientation Error ((\Theta))</th>
<th>Position Error (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUM VI Room 1 (orig)</td>
<td>1.451</td>
<td>0.056</td>
</tr>
<tr>
<td>TUM VI Room 1 (int)</td>
<td>1.854</td>
<td>0.054</td>
</tr>
<tr>
<td>TUM VI Room 2 (orig)</td>
<td>1.448</td>
<td>0.071</td>
</tr>
<tr>
<td>TUM VI Room 2 (int)</td>
<td>1.422</td>
<td>0.074</td>
</tr>
<tr>
<td>TUM VI Room 3 (orig)</td>
<td>1.354</td>
<td>0.061</td>
</tr>
<tr>
<td>TUM VI Room 3 (int)</td>
<td>1.428</td>
<td>0.076</td>
</tr>
</tbody>
</table>
FPGA Results

For a $512 \times 512$ input image

<table>
<thead>
<tr>
<th></th>
<th>Vision</th>
<th>IMU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock Period</td>
<td>12 ns</td>
<td>8 ns</td>
</tr>
<tr>
<td>BRAM</td>
<td>243 (12%)</td>
<td>16 (~0%)</td>
</tr>
<tr>
<td>DSP</td>
<td>1559 (79%)</td>
<td>756 (38%)</td>
</tr>
<tr>
<td>FF</td>
<td>230272 (12%)</td>
<td>64032 (3%)</td>
</tr>
<tr>
<td>LUT</td>
<td>553169 (61%)</td>
<td>99706 (11%)</td>
</tr>
<tr>
<td>URAM</td>
<td>21 (4%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>
Visual Pipeline Preliminary Latency per Kernel

For a $512 \times 512$ input image

✓ Functionally correct
× Optimization in progress

<table>
<thead>
<tr>
<th>Function</th>
<th>Clock (ns)</th>
<th>Cycles</th>
<th>Latency (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histogram Equalization</td>
<td>5</td>
<td>771630</td>
<td>3.86</td>
</tr>
<tr>
<td>FAST corner detection</td>
<td>12</td>
<td>1202358</td>
<td>14.4</td>
</tr>
<tr>
<td>Pyramid Generation + Optical Flow</td>
<td>12</td>
<td>10587926</td>
<td>127</td>
</tr>
<tr>
<td>Undistortion</td>
<td>5</td>
<td>43249</td>
<td>0.216</td>
</tr>
<tr>
<td>RANSAC</td>
<td>5</td>
<td>2915580</td>
<td>14.6</td>
</tr>
<tr>
<td>Estimated Total</td>
<td></td>
<td></td>
<td>$\sim 290$</td>
</tr>
</tbody>
</table>
IMU Pipeline Preliminary Latency

Batching 11 IMU measurements together
✓ Functionally correct
✓ Optimized

<table>
<thead>
<tr>
<th></th>
<th>Clock (ns)</th>
<th>Cycles</th>
<th>Latency (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMU pipeline</td>
<td>5.3</td>
<td>1285844</td>
<td>6.82</td>
</tr>
</tbody>
</table>
1. Continue to improve on FPGA prototype
2. Develop and tapeout SoC for accelerating visual-inertial odometry and rest of perception pipeline in November
3. Evaluate performance/power/quality benefit of accelerator in context of entire XR application pipeline
4. Research further into the other pipelines (visual and audio) and integrate with the perception pipeline accelerator