Evaluating the Scalability of Stencil Codes at Scale

*Manish Modani, Rupert Ford
Daresbury Laboratory, UK

Constantinos Evangelinos
IBM Research, USA

(EASC 22/4/2015)
Outline

- Introduction
- Benchmark
- Observation/Issues
- Proposed Mapping
- Energy Measurement
- Conclusions
Communication at Scale

- We have witnessed an ever slower increase and, most recently, reduction in the clock speeds of microprocessor cores.

- This has resulted in higher core counts – the largest systems today have in excess of 100,000 cores and this will increase further for exascale (systems in 1995 had around 512 microprocessors).

- The designing of parallel applications needs to be revisited at all the stages for higher core counts (e.g. Partitioning / Communication / Mapping).

- Communication cost with neighbours is the key for the performance.
Benchmark

• Stencil-based codes are widely used in Scientific Computing and are considered to be good candidates for running at scale.

• 2D stencil kernel has been written to test the assumption that stencil-based codes scale.

• This kernel performs the halo communication that a stencil code would require but does no computation.

• The halo communication is repeated large number of times.
A (N,E,S,W) communication pattern is used which is representative of a 2-dimensional partitioning of a 5-point stencil on regular grid.
Weak Scaling

- Considered weak scaling to have better idea about the code's scalability.
- Each task communicates with the same number of neighbour's and communicates the same amount of data, irrespective of the number of tasks being used.
- As the computation per task remains the same, the compute to communicate ratio remains the same.
MPI IMPLEMENTATION

• Communication using MPI.

• MPI Dead-Lock:
  – Even & Odd ranks communication.

• This kernel run using a range of problem sizes on a Blue Gene Q up to 65,000 cores and on TITAN & ARCHER up to 16,000 cores.
Message Size

The halo sizes range from 3,200 bytes per halo (100 levels * 4 columns * 8) to 204,800 bytes per halo (100 levels * 256 columns * 8) bytes.

The particular pattern and sizes were chosen as they cover what is used by a large number of Atmosphere models in Climate and Weather Forecasting however the results are relevant to other disciplines.
Blue Joule - Blue Gene/Q:
Hartree Centre UK
Blue Joule consists of 6 racks,
Each rack containing 1,024 nodes
Each node has a 16-core, 64 bit
A2 Power PC, 1.60 GHz processor.
Nodes are interconnected by 5D TORUS

Not Scaling
Observation: Titan

Titan: CRAY XK7

Titan: Oak Ridge National Laboratory, USA.
CRAY XK7
18,688 AMD Opteron Cores
16-core CPUs
Gemini Interconnect

Not Scaling
Observation ARCHER

ARCHER UK:
Cray XC30-Architecture,
2.7 GHz Ivy Bridge,
24 cores per node,
Aries Interconnect-
Dragonfly topology
Nodes are connected
to each Aries router;
188 nodes are
grouped into a cabinet;
and two cabinets make
up a group

Not Scaling
BGQ Tool: Cartesian Topology

Cartesian-Topology Utility to map MPI Ranks on TORUS Network
Requires information about the shape of the block.
Cray performance tool
Perftools Module: Using Pat_build and Pat_report a mpi rank file generated. Rerun the code with new rank file using the variable MPICH_RANK_REORDER_METHOD=3
Analysis

The performance of the code was analysed for 1 dimensional.

It was observed that the communication is X directions scales well.

In Y direction it does not, where most of the messages travels outside the node.

For Y direction, message needed to travel many hops.
5DTORUS : BGQ
Node Board (32 Compute Nodes): 2x2x2x2x2

Rank (A,B,C,D,E)
Rank 29 (1,0,0,0,0)
Rank 28 (1,0,1,0,0)
Rank 31 (1,1,1,0,0)

Rank 17 (1,0,0,1,0)
Rank 20 (1,1,1,1,1)
Task-to-Topology Mapping: Intra-Node Communication

- Default Layout: 16x16: 256 MPI task

<table>
<thead>
<tr>
<th>240</th>
<th>241</th>
<th>242</th>
<th>243</th>
<th>244</th>
<th>245</th>
<th>246</th>
<th>247</th>
<th>248</th>
<th>249</th>
<th>250</th>
<th>251</th>
<th>252</th>
<th>253</th>
<th>254</th>
<th>255</th>
</tr>
</thead>
<tbody>
<tr>
<td>224</td>
<td>225</td>
<td>226</td>
<td>227</td>
<td>228</td>
<td>229</td>
<td>230</td>
<td>231</td>
<td>232</td>
<td>233</td>
<td>234</td>
<td>235</td>
<td>236</td>
<td>237</td>
<td>238</td>
<td>239</td>
</tr>
<tr>
<td>208</td>
<td>209</td>
<td>210</td>
<td>211</td>
<td>212</td>
<td>213</td>
<td>214</td>
<td>215</td>
<td>216</td>
<td>217</td>
<td>218</td>
<td>219</td>
<td>220</td>
<td>221</td>
<td>222</td>
<td>223</td>
</tr>
<tr>
<td>192</td>
<td>193</td>
<td>194</td>
<td>195</td>
<td>196</td>
<td>197</td>
<td>198</td>
<td>199</td>
<td>200</td>
<td>201</td>
<td>202</td>
<td>203</td>
<td>204</td>
<td>205</td>
<td>206</td>
<td>207</td>
</tr>
<tr>
<td>176</td>
<td>177</td>
<td>178</td>
<td>179</td>
<td>180</td>
<td>181</td>
<td>182</td>
<td>183</td>
<td>184</td>
<td>185</td>
<td>186</td>
<td>187</td>
<td>188</td>
<td>189</td>
<td>190</td>
<td>191</td>
</tr>
<tr>
<td>160</td>
<td>161</td>
<td>162</td>
<td>163</td>
<td>164</td>
<td>165</td>
<td>166</td>
<td>167</td>
<td>168</td>
<td>169</td>
<td>170</td>
<td>171</td>
<td>172</td>
<td>173</td>
<td>174</td>
<td>175</td>
</tr>
<tr>
<td>144</td>
<td>145</td>
<td>146</td>
<td>147</td>
<td>148</td>
<td>149</td>
<td>150</td>
<td>151</td>
<td>152</td>
<td>153</td>
<td>154</td>
<td>155</td>
<td>156</td>
<td>157</td>
<td>158</td>
<td>159</td>
</tr>
<tr>
<td>128</td>
<td>129</td>
<td>130</td>
<td>131</td>
<td>132</td>
<td>133</td>
<td>134</td>
<td>135</td>
<td>136</td>
<td>137</td>
<td>138</td>
<td>139</td>
<td>140</td>
<td>141</td>
<td>142</td>
<td>143</td>
</tr>
<tr>
<td>112</td>
<td>113</td>
<td>114</td>
<td>115</td>
<td>116</td>
<td>117</td>
<td>118</td>
<td>119</td>
<td>120</td>
<td>121</td>
<td>122</td>
<td>123</td>
<td>124</td>
<td>125</td>
<td>126</td>
<td>127</td>
</tr>
<tr>
<td>96</td>
<td>97</td>
<td>98</td>
<td>99</td>
<td>100</td>
<td>101</td>
<td>102</td>
<td>103</td>
<td>104</td>
<td>105</td>
<td>106</td>
<td>107</td>
<td>108</td>
<td>109</td>
<td>110</td>
<td>111</td>
</tr>
<tr>
<td>80</td>
<td>81</td>
<td>82</td>
<td>83</td>
<td>84</td>
<td>85</td>
<td>86</td>
<td>87</td>
<td>88</td>
<td>89</td>
<td>90</td>
<td>91</td>
<td>92</td>
<td>93</td>
<td>94</td>
<td>95</td>
</tr>
<tr>
<td>64</td>
<td>65</td>
<td>66</td>
<td>67</td>
<td>68</td>
<td>69</td>
<td>70</td>
<td>71</td>
<td>72</td>
<td>73</td>
<td>74</td>
<td>75</td>
<td>76</td>
<td>77</td>
<td>78</td>
<td>79</td>
</tr>
<tr>
<td>48</td>
<td>49</td>
<td>50</td>
<td>51</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>55</td>
<td>56</td>
<td>57</td>
<td>58</td>
<td>59</td>
<td>60</td>
<td>61</td>
<td>62</td>
<td>63</td>
</tr>
<tr>
<td>32</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>36</td>
<td>37</td>
<td>38</td>
<td>39</td>
<td>40</td>
<td>41</td>
<td>42</td>
<td>43</td>
<td>44</td>
<td>45</td>
<td>46</td>
<td>47</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
</tbody>
</table>

Default: 32 communication outside node
**Intra-Node Communication**

- Minimum communication outside a node

<table>
<thead>
<tr>
<th>204</th>
<th>205</th>
<th>206</th>
<th>207</th>
<th>220</th>
<th>221</th>
<th>222</th>
<th>223</th>
<th>236</th>
<th>237</th>
<th>238</th>
<th>239</th>
<th>252</th>
<th>253</th>
<th>254</th>
<th>255</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>201</td>
<td>202</td>
<td>203</td>
<td>216</td>
<td>217</td>
<td>218</td>
<td>219</td>
<td>232</td>
<td>233</td>
<td>234</td>
<td>235</td>
<td>248</td>
<td>249</td>
<td>250</td>
<td>251</td>
</tr>
<tr>
<td>196</td>
<td>197</td>
<td>198</td>
<td>199</td>
<td>212</td>
<td>213</td>
<td>214</td>
<td>215</td>
<td>228</td>
<td>229</td>
<td>230</td>
<td>231</td>
<td>244</td>
<td>245</td>
<td>246</td>
<td>247</td>
</tr>
<tr>
<td>192</td>
<td>193</td>
<td>194</td>
<td>195</td>
<td>208</td>
<td>209</td>
<td>210</td>
<td>211</td>
<td>224</td>
<td>225</td>
<td>226</td>
<td>227</td>
<td>240</td>
<td>241</td>
<td>242</td>
<td>243</td>
</tr>
<tr>
<td>140</td>
<td>141</td>
<td>142</td>
<td>143</td>
<td>156</td>
<td>157</td>
<td>158</td>
<td>159</td>
<td>172</td>
<td>173</td>
<td>174</td>
<td>175</td>
<td>188</td>
<td>189</td>
<td>190</td>
<td>191</td>
</tr>
<tr>
<td>136</td>
<td>137</td>
<td>138</td>
<td>139</td>
<td>152</td>
<td>153</td>
<td>154</td>
<td>155</td>
<td>168</td>
<td>169</td>
<td>170</td>
<td>171</td>
<td>184</td>
<td>185</td>
<td>186</td>
<td>187</td>
</tr>
<tr>
<td>132</td>
<td>133</td>
<td>134</td>
<td>135</td>
<td>148</td>
<td>149</td>
<td>150</td>
<td>151</td>
<td>164</td>
<td>165</td>
<td>166</td>
<td>167</td>
<td>180</td>
<td>181</td>
<td>182</td>
<td>183</td>
</tr>
<tr>
<td>128</td>
<td>129</td>
<td>130</td>
<td>131</td>
<td>144</td>
<td>145</td>
<td>146</td>
<td>147</td>
<td>160</td>
<td>161</td>
<td>162</td>
<td>163</td>
<td>176</td>
<td>177</td>
<td>178</td>
<td>179</td>
</tr>
<tr>
<td>76</td>
<td>77</td>
<td>78</td>
<td>79</td>
<td>92</td>
<td>93</td>
<td>94</td>
<td>95</td>
<td>108</td>
<td>109</td>
<td>110</td>
<td>111</td>
<td>124</td>
<td>125</td>
<td>126</td>
<td>127</td>
</tr>
<tr>
<td>72</td>
<td>73</td>
<td>74</td>
<td>75</td>
<td>88</td>
<td>89</td>
<td>90</td>
<td>91</td>
<td>104</td>
<td>105</td>
<td>106</td>
<td>107</td>
<td>120</td>
<td>121</td>
<td>122</td>
<td>123</td>
</tr>
<tr>
<td>68</td>
<td>69</td>
<td>70</td>
<td>71</td>
<td>84</td>
<td>85</td>
<td>86</td>
<td>87</td>
<td>100</td>
<td>101</td>
<td>102</td>
<td>103</td>
<td>116</td>
<td>117</td>
<td>118</td>
<td>119</td>
</tr>
<tr>
<td>64</td>
<td>65</td>
<td>66</td>
<td>67</td>
<td>80</td>
<td>81</td>
<td>82</td>
<td>83</td>
<td>96</td>
<td>97</td>
<td>98</td>
<td>99</td>
<td>112</td>
<td>113</td>
<td>114</td>
<td>115</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td>44</td>
<td>45</td>
<td>46</td>
<td>47</td>
<td>60</td>
<td>61</td>
<td>62</td>
<td>63</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>40</td>
<td>41</td>
<td>42</td>
<td>43</td>
<td>56</td>
<td>57</td>
<td>58</td>
<td>59</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>36</td>
<td>37</td>
<td>38</td>
<td>39</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>55</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>32</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>48</td>
<td>49</td>
<td>50</td>
<td>51</td>
</tr>
</tbody>
</table>

Proposed: 16 communication outside node
Task-to-Topology Mapping

- MPI Ranks re-ordered to travel $\leq 1$ hop for 4x4x4x8x2 (128x128; 16384 mpi tasks)

Default

Proposed
Task-to-Topology Mapping

- MPI Ranks re-ordered to travels $\leq 1$ hop e.g. for 4x4x4x8x2 (128x128; 16384 mpi tasks)

```
        11439 (2,3,0,5,0)
         /    \
        4     1
   11311 (2,3,0,1,0) 11312 (2,3,0,1,1)

11183 (2,3,3,5,0)
```

Default

```
        11439 (1,1,2,5,0)
         /    \
        0     0
   11310 (1,1,2,5,0) 11312 (1,1,2,5,0)

11183 (2,1,2,5,0)
```

Proposed
Results

BGQ: 1 hop Mapping

Execute Time (Secs)

Processors
Energy Measurement: BGQ

- EMONSimple is a simple energy monitoring library for Blue Gene/Q. It provides a trace of power consumption versus time for an executable.

- Each BG/Q Node-Board (collection of 32 BG/Q nodes) contains an FPGA that records instantaneous power consumption.

- The sampling frequency of the FPGA is ~ 0.3s and the sampled information can be read by a program via the BG/Q's EMON API.

- This energy information is output at the end of the program giving a trace of the programs energy consumption.

- Easy to use: Link EMON library to binary.
BGQ-Energy: Default
BGQ-Energy: Mapping

Integrated Power - Node Board

Power (W)

Time (s)
Conclusions

- Scalability for weak scaling for 2D-communication to neighbours at scale were observed at scale.
- The default and system tools suggested rank placement does not benefit.
- All the messages shares the same link hence the contention for link bandwidth is the issue.
- 1hop task-to-toppology mapping scheme devised.
- Energy measurement shows upto 60 % less energy consumption from the devised mapping.
Future Direction

Analysis/New Mapping for 3D stencil code.

Acknowledgments

We acknowledge the support provided by Hartree, ORNL & ARCHER Team to use their HPC systems.
Thanking you