Intelligent On/Off Link Management for On-Chip Networks

Andreas G. Savva
University of Cyprus
eep7sa4@ucy.ac.cy

Theocharis Theocharides
University of Cyprus
ttheocharides@ucy.ac.cy

Vassos Soteriou
Cyprus University of Technology
vassos.soteriou@cut.ac.cy

ABSTRACT

Links connecting on-chip components are a major source of power consumption in modern-day on-chip interconnects. Several efforts have henceforth focused on reducing the power consumption, the majority of which target selected links for turning on and off. In this paper, we propose an intelligent power management policy for networks-on-chip where links are turned off and switched back on based on a neural network, which processes link utilization as feedback from the system and determines which links are candidates for turning off and back on. The neural network keeps relatively small in terms of area and power consumption, as it is used to forecast the optimal utilization threshold for which underutilized links are turned off.

1. INTRODUCTION

Power consumption has been identified as one of the major issues affecting networks-on-chip (NoCs), and a relatively large effort has been undertaken in both academia and industry related to addressing power issues surrounding these systems [1, 2, 4, 5, 6]. Most of the existing technologies focus on DVFS (Dynamic Voltage and Frequency Scaling) or turning links on/off via a statically determined threshold.

This paper introduces an Artificial Neural Network (ANN) methodology for managing power consumption in NoCs, based on the idea of dynamically turning on/off links, in response to the existing communication traffic information. An integrated hardware-based ANN is used in order to predict which links should be turned off and on, based on an intelligently and dynamically computed threshold. The ANN receives link utilization data in discrete time intervals, and based on its training, it outputs an intelligently computed threshold for which links are turned on and off for the next discrete interval. In contrast to forecasting the actual links that would be turned on and off, computing only the threshold maintains the size of the ANN hardware relatively small, keeping the ANN power consumption to the minimum and thus positively impacting the overall power savings. In addition, the ANN can be designed hierarchically and can easily be reused and scaled to whatever NoC size necessary. Simulation results in 2-D Mesh and Torus topologies using synthetic traffic models, show an average power savings of approximately 13% when compared to non-intelligent techniques and also better savings when compared to related works.

2. PROPOSED METHODOLOGY

We partition the NoC into smaller (i.e., 4×4 regions), and assign a base ANN architecture to monitor each region. Therefore, we limit the number of inputs to the ANN, restricting the ANN size to a minimum. The ANN-based mechanism can be integrated as an independent PE and each base ANN mechanism can be assigned to monitor an NoC partition. Both cases are highlighted in an example setup, in Figure 1.

Figure 1: ANN predictor with NoCs and an 8×8 network partition into four 4×4 networks with their ANNs.

The ANN mechanism receives the link utilization of all the routers of the NoC partition that is responsible for monitoring. Each router keeps a counter that keeps track of the packets travelling on each link, and sends the counter value to the ANN every n cycles. The ANN mechanism computes the minimum values of all the link utilizations during each interval, and feeds them to the ANN. If a router fails to transmit the values during a single interval, its value is set to a sentinel value, indicating that the buffers of that router are fully utilized, as they are potentially blocked. This inherently built time-out mechanism serves also as a congestion information mechanism, as links involved in heavily congested routers are not likely candidates to be turned off, as they are likely to be heavily active. The ANN then uses these utilization values, to compute the new threshold that is then used to determine whether a link is going to be turned off or turned back on, for the next n-cycle interval. An overview of the procedure for a 4×4 NoC partition is shown in Figure 2.

Figure 2: Main steps of a 4×4 ANN predictor.
3. SIMULATION AND RESULTS

We compared the power savings as well as the throughput of the dynamic, ANN-based on/off link prediction algorithm, to a static threshold-based algorithm. Furthermore, we also compared the algorithms to a system without any on/off mechanism. We adopted the Orion power model for the dynamic power consumption of each router [7], and designed and synthesized the router and link hardware in Verilog and Synopsys Design Compiler, in an attempt to obtain the leakage power. Figure 3 summarizes the comparison when targeting a 4x4 mesh NoC, and 8x8 mesh and torus NoCs. For all traffic models, the power savings of the ANN-based mechanism outperform both the statically-determined case, and the case without any on/off links.

Next, we measure the performance implications associated with the ANN threshold computation. Figure 4 shows the overall network bandwidth, which involves all data (payload+control) that travels on the network. Obviously, the ANN-based technique yields better results when compared to the statically-determined threshold. We also evaluate the hardware overheads of the proposed scheme. The ANN-based mechanism was synthesized and implemented targeting a commercial 65nm CMOS technology. The ensuing synthesized ANN-based controller and the associated hardware overheads in each router, consume approximately 4K logic gates (for comparison purposes, an NoC router similar to the one used in our simulation consumes roughly 21K gates), bringing the estimated hardware overhead for an 4x4 mesh network to roughly 4% of the NoC hardware.

Lastly, we briefly give a comparison with relevant related works in Table 1. When compared to both [1] and [3], the ANN-based prediction yields better power savings than having no prediction mechanism, while still maintaining lower hardware overheads.

<table>
<thead>
<tr>
<th>Related Work</th>
<th>Characteristics</th>
<th>Power savings comparing to no algorithm</th>
<th>Hardware Overhead</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>8x8 2-D mesh topology, Uniform traffic</td>
<td>~ 37.5% - turning on/off 2 links per router</td>
<td>N/A</td>
</tr>
<tr>
<td>[3]</td>
<td>8x8 2-D mesh topology, Pareto distribution – 0.5 packet injection rate</td>
<td>~ 30%</td>
<td>500 equivalent logic gates per router port</td>
</tr>
<tr>
<td>Proposed ANN-based Technique</td>
<td>8x8 2-D mesh topology and 8x8 torus topology</td>
<td>Up to ~ 40% - based on ANN prediction</td>
<td>4% of the NoC hardware for a complete 4x4 Mesh NoC</td>
</tr>
</tbody>
</table>

Table 1: Power savings/hardware overhead comparisons.

4. CONCLUSIONS AND FUTURE WORK

This paper presented how an ANN-based mechanism can be used to dynamically compute a utilization threshold, that can in turn be used to select candidate links for turning on or off, in an effort to achieve power savings in an NoC. Future work includes optimization of the routing algorithm to take into consideration the ANN-based computation, as well as exploitation of the results in real application traffic models. We believe that runtime, intelligent algorithms for power management can be very beneficial, especially in large and complex systems such as NoCs.

REFERENCES