Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

OPTIMIZING PRODUCT MIX PROFILE FOR MAXIMUM OUTPUT AND STABLE LINE PERFOMANCE IN A GIGA FAB

Georg Seidel

Infineon Technologies Austria AG
Siemensstraße 2
Villach 9500, AUSTRIA

Chien Yong Low
Mun Hoe Hooi
Soo Leen Low
Boon Ping Gan

D-SIMLAB Technologies Pte Ltd
8 Jurong Town Hall Road #23-05 JTC Summit
Singapore 609434, SINGAPORE

Birgit Hoelscher

Infineon Technologies GmbH
Am Campeon 1-15
Neubiberg 85579, GERMANY

ABSTRACT

Chip manufacturers are trying recently to maximize chip output with their existing capacity due to the worldwide chip shortage, because expanding capacity within a short time span is not a practical option. But many times maximizing chip output results in unstable production line performance, and creates cyclical actions of increasing and cutting production volume, which does not necessarily lead to maximum output over a period of time. In this paper, we present an approach that combines a greedy optimization algorithm, a static capacity model, and a dynamic simulation model to maximize production output, and ensuring stable production line performance. Our approach managed to achieve an 11% output increase as compared to a manual planning effort, and at the same time, achieved a stable dynamic flow factor.

1 INTRODUCTION

When the Covid-19 pandemic started in early 2020, manufacturers significantly reduced production due to the anticipated reduction in chip demand. In contrary, the chip demand increased during the lockdown period. Due to Covid restrictions, supply chain issues, and generally a long lead time to increase production capacity the manufacturers were not able to ramp production quickly. These created a huge demand backlog, and a worldwide chip shortage. One of the key bottlenecks in the chip production supply chain is the wafer fabrication plant. Wafer fabs are thus coping with this issue by attempting to maximize production output with the existing capacity. This often results in an unstable fab performance as the complex interaction between different product capacity corridors are not obvious. To solve this problem, we devised an approach of combining a greedy optimization algorithm, a static capacity model and a dynamic simulation to achieve maximum production output and ensuring a stable line performance. The approach was tested for Infineon's fab in Kulim, Malaysia, which has a monthly output of more than 120,000 wafers.

2 LOAD MIX OPTIMIZATION AND RESULTS

The Load Mix Optimizer (LMO) Engine initiates its first iteration run by feeding the D-SIMCON Static Capacity Engine (D-SCE) with the initial product mix profile that was provided by Infineon. The D-SCE calculates the expected utilization for each work center with the provided mix profile by balancing utilization of equipment within the work center. The resulting work center utilization is then compared with the utilization limit of the work center. The utilization limit is set for each work center independently. If any of the work center utilization is above the utilization limit, the LMO reduces the volume for products that consume the most bottleneck work center capacity. On the other hand, if all work center utilizations are below the utilization limits, the LMO will increase the volume for products that consumes the least bottleneck work center capacity. Selecting products in this way is greedy in nature. Each product volume is bounded by a minimum and maximum value. Once a new product mix profile is determined, the LMO triggers the D-SCE to re-calculate the expected work center utilization. This process iterates until the improvement of the defined objective function remains lower than a defined threshold, *t*, for x consecutive iterations, where t and x are configurable parameters. The objective function in our case is maximum production loading. The optimized load mix is then fed to a dynamic simulation model built on D-SIMCON Simulation Engine to ensure the dynamic flow factor is within a defined range. If the dynamic flow factor increases too much over time, the optimized load mix will be rejected, and the LMO run is re-triggered to find an alternative product mix profile.

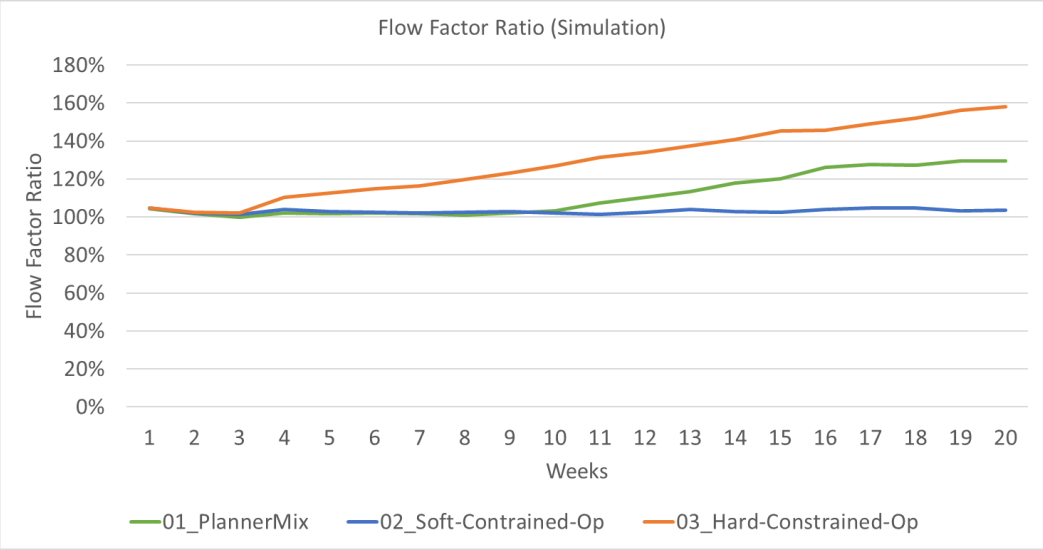

Figure 1: Optimized product mix gained and effect on flow factor.

Using the LMO algorithm, two optimized product mix profiles were generated. The first profile (soft constrained profile) was generated by setting the work center utilization limits below the expected uptimes, while the second profile (hard constrained profile) was generated using the expected uptimes as utilization limits. With the first "soft" profile, we observed a gain of 11% in layer starts per week (LSPW) relative to the planner mix profile. While in the second "hard" profile, we observed a gain of 33% in LPSW. A simulation was run with these three mix profiles, and the flow factor ratio (calculated as flow factor of the week divided by minimum flow factor across the weeks) are compared in Figure 1. The "hard" profile created an unstable production run as the flow factor ratio increased 150% over time (the orange line). The "soft" profile provides a stable flow factor (blue line), which signifies it's feasibility for production run. One key point to take note of is that the planner mix profile resulted in an unstable production line too, with the flow factor ratio increases 120% (green line). The strength of the LMO algorithm shows tremendous potential to be adopted for production planning as it manages to get a higher output, without compromising on production line performance.