3D-Floorplanning für hochparallele Verbindungsstrukturen

Johann Knechtel, Jens Lienig, Sergii Osmolovskyi

johann.knechtel@ifte.de, jens.lienig@ifte.de, sergii.osmolovskyi@ifte.de
Institut für Feinwerktechnik und Elektronik-Design, Technische Universität Dresden

KURZFASSUNG

Dreidimensional integrierte Schaltkreise (3D-ICs) beruhen auf hochparallelen Verbindungsstrukturen, z.B. Bussen, um die Kommunikation zwischen den verteilten Blöcken eines 3D-ICs zu gewährleisten. Durch die gezielte Ausrichtung von Blöcken während des Floorplanning lassen sich derartige Verbindungsstrukturen effektiv planen und realisieren. Eine Erweiterung der bewährten Corner Block List, effiziente Algorithmen für die Layouterstellung, sowie eine zielgerichtete Optimierung (basierend auf der Simulated-Annealing-Heuristik) sind drei wesentliche Merkmale der neuen 3D-Floorplanning-Methodik. Neben der Ausrichtung von Blöcken berücksichtigt sie wichtige Kriterien des 3D-IC-Entwurfs, wie das thermische Management oder die Einhaltung von fixen Chipabmessungen. Experimentelle Untersuchungen anhand der GSRC-Benchmarks zeigen das Potential unseres Ansatzes sowohl für die Planung von hochparallelen Verbindungsstrukturen als auch für das 3D-Floorplanning.

I. EINLEITUNG

Das dreidimensionale Stapeln von Chips zu sog. 3D-integrierten Schaltkreisen (3D-ICs) ist ein vielversprechender Ansatz, um den heutigen und zukünftigen Anforderungen an Leistung, Funktionalität und Stromaufnahme von ICs gerecht zu werden. Dank den vertikalen Verbindungen eines 3D-ICs zwischen dessen einzelnen Chips, vor allem den Through-Silicon Vias (TSVs), sind kurze Verdrahtung und damit leistungsfähige ICs realisierbar, wie z.B. in [1], [2] demonstriert.

Der Ansatz der *gezielten Blockausrichtung* für die Planung von *hochparallelen Verbindungsstrukturen* bzw. *Bussen* findet im klassischen 2D-Floorplanning erfolgreich Anwendung [3], [4]. Dabei besteht die Zielstellung darin, Blöcke derart auszurichten, dass die dem Floorplanning nachfolgende Verdrahtung alle beteiligten Bus-Signale in parallelen, kürzestmöglichen Leiterzügen realisieren kann. Für das 3D-Floorplanning wurden bisher kaum derartige Ansätze vorgestellt. Insbesondere bei der blockbasierten 3D-Integration als der bekanntesten Entwurfsmethodik [5], [6] ist die gezielte Ausrichtung bei der Planung von Bussen relevant. Wie in Abb. 1 erläutert, erlauben verschiedene, spezifische Blockausrichtungen die Berücksichtigung aller wesentlicher Bus-Varianten innerhalb eines 3D-ICs.

II. CORBLIVAR: CORNER BLOCK LIST FÜR VARIABLE (BLOCK-)AUSRICHTUNG

Für die gezielte Ausrichtung von Blöcken zur Planung von Bussen wird nachfolgend eine Erweiterung der klassischen, zweidimensionalen Corner Block List (CBL) [7] vorgestellt, genannt *Corblivar*. Ein 3D-IC mit n Chips wird in Corblivar mittels einer Sequenz $\{CBL_1, \ldots, CBL_n\}$ von klassischen CBL-Tupeln, sowie einer neuartigen globalen *Ausrichtungssequenz* A kodiert.

Die *Tupel zur paarweisen Blockausrichtung* $a_k \in A = \{a_1, \ldots, a_m\}$ erlauben eine Kodierung aller in Abb. 1 beschriebenen Verbindungsstrukturen. Die Tupel sind jeweils definiert als $a_k = (b_i, b_j, (AR_x, ART_x), (AR_y, ART_y))$, wobei b_i und b_j zwei zur Planung von Verbindungsstrukturen auszurichtende Blöcke beschreibt, sowie (AR_x, ART_x) und (AR_y, ART_y) jeweils die Ausrichtungen bzgl. der x- und y-Koordinaten beschreiben. Diese Kodierung erlaubt damit eine *unabhängige und gleichzeitige* Ausrichtung von Blöcken in x- und y-Richtung. Damit können alle verschiedenen Arten der in Abb. 1 beschriebenen Verbindungsstrukturen modelliert werden. Die entsprechenden Varianten der Blockausrichtung sind definiert als *fixer Offset* ($ART = 0$), *minimale Überschneidung* ($ART = 1$), *maximaler Abstand* ($ART = 2$) und *don't care* ($ART = -1$). Zum Beispiel kann der klassische Bus C in Abb. 1 mittels zweier Tupel beschrieben werden; ein Tupel definiert die paarweise Ausrichtung des linken und mittleren Blockes, und ein zweites Tupel definiert die Ausrichtung des linken und rechten

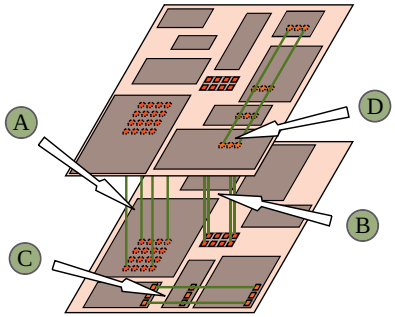

Abbildung 1. Hochparallele Verbindungsstrukturen innerhalb eines 3D-ICs und deren entsprechenden Blockausrichtungen. *Vertikale Busse* (A) verbinden auf mehrere Chips verteilte Blöcke. *TSV-Stacks* (B), bestehend aus einem Bündel ausgerichteter TSVs, finden beispielsweise Anwendung in Network-on-Chip (NoC)-Strukturen. Beide Strukturen basieren auf flexibler bzw. fixer *Inter-Chip-Ausrichtung*, d.h. die Blöcke bzw. TSVs sind zwischen verschiedenen Chips auszurichten. *Klassische 2D-Busse*, mit fixen oder flexiblen Anschlusspins (C oder D), werden z.B. für Datenpfad-basierte Komponenten genutzt. Die Busse setzen eine Ausrichtung innerhalb eines Chips voraus (*Intra-Chip-Ausrichtung*).

Blockes. Die Ausrichtungen in x-Richtung basieren dabei auf maximalen Abständen, um die Länge des Busses zu begrenzen. Die Ausrichtungen in y-Richtung nutzen einen fixen Offset (punktgenaue Ausrichtung), um für die festen Anschlusspins eine entsprechend geradlinig horizontale Verdrahtung zu ermöglichen.

III. 3D-FLOORPLANNING-METHODIK UND IMPLEMENTIERUNG

Den Aufbau der vorgestellten 3D-Floorplanning-Methodik stellt Abb. 2 vor. Sie erweitert die Ansätze der CBL derart, dass gleichzeitig Blockausrichtungen sowohl zwischen (Intra-) als auch innerhalb (Inter-Chip) von Chips realisierbar sind und die Ausrichtungen fixe und flexible Abstände berücksichtigen können. Um diese verschiedenen Anforderungen zu realisieren, ist der Prozess zur Layouterstellung in einer hierarchischen und "orchestrierenden" Art und Weise implementiert. Dabei überwacht ein zentraler Prozess sowohl die Layouterstellung innerhalb aller einzelner Chips, als auch die verschiedenen Blockausrichtungen. Dieser zentrale Prozess delegiert entsprechend an lokale Prozesse zur Platzierung als auch zur gezielten Ausrichtung von Blöcken.

Der Optimierungsprozess der 3D-Floorplanning-Methodik basiert auf der bekannten Simulated-Annealing (SA)-Heuristik. Es hat sich bei unseren Untersuchungen gezeigt, dass bisherige Ansätze des SA-basierten 3D-Floorplanning bei der Exploration des Lösungsraums, vor allem bei Anwendung von gezielten Blockausrichtungen, nicht effektiv genug sind. Daher beinhaltet unsere Methodik einen robusten, da adaptiven Optimierungsablauf. Dieser zeichnet sich dadurch aus, dass die Kostenänderungen während des Optimierungsprozesses kontinuierlich geprüft werden. Eine Stagnation der Kosten für eine gewisse Zeit deutet darauf hin, dass der Prozess in ein lokales Minima geraten ist. In solchen Situationen wird die SA-Temperatur schrittweise erhöht, bis sich eine deutlich höhere Variation der Kosten einstellt, d.h., bis der Prozess dem lokalen Minima "entfliehen" kann. Dieser adaptive Optimierungsablauf, in Kombination mit einer in späten Optimierungsphasen gezielten Umplatzierung von (bis dahin fehlerhaft) ausgerichteten Blöcken, leistet laut unseren Experimenten einen wesentlichen Beitrag zur erfolgreichen Planung einer Vielzahl von verschiedenartigen Busstrukturen.

Die schnelle und gleichzeitig akkurate thermische Analyse der 3D-Floorplanning-Methodik erweitert das Prinzip des *Power Blurring* [8]. Dies basiert auf, im Vergleich zu Finite-Elemente-Analysen (FEA) simplen, mathematischen Faltungen von Leistungsdichten und sogenannten thermischen Impulsantworten. Solche Impulsantworten modellieren die Wärmeleitung innerhalb einzelner Chips, ausgehend von einer gedachten Punktquelle. Im Gegensatz zu [8] wird

J. Knechtel, J. Lienig, S. Osmolovskyi "3D-Floorplanning für hochparallele Verbindungsstrukturen," Tagungsband Dresdner Arbeitstagung Schaltungs- und Systementwurf (DASS 2014), Fraunhofer Verlag, ISBN 978-3-8396-0738-1, pp. 16-20, April 2014.

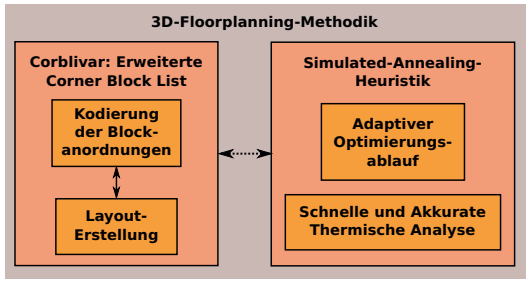

Abbildung 2. Prinzipieller Aufbau der 3D-Floorplanning-Methodik.

hier keine FEA angewandt, um die Impulsantworten zu generieren. Stattdessen sind die Impulsantworten mittels symmetrischen, zweidimensionalen Gauss-Funktionen $g(x, y, w, s) = w \exp\left(-\frac{1}{s}x^2\right) \exp\left(-\frac{1}{s}y^2\right)$ beschrieben, wobei w den Amplitudenfaktor und s den Streuungsfaktor darstellt. Die für jeden einzelnen Chip eines 3D-ICs benötigten verschiedenen Impulsantworten werden mittels Skalierung generiert. Konkret ist w für jeden Chip d_i derart zu skalieren, dass $w_i = w/(i^{w_s})$ den jeweiligen Amplitudenfaktor beschreibt. Dabei gilt, dass $d_{\max(i)}$ den obersten Chip beschreibt, welcher per Annahmen direkt mit dem Heatsink verbunden ist. Die eigentliche Parametrisierung von w, w_s und s erfolgt für jede Konfiguration eines 3D-ICs, d.h., für jede Variation von Anzahl der Chips und deren Abmessungen. Dazu wird zuerst eine thermische Verteilung mittels *HotSpot* [9] generiert, welche dann als Referenzlösung gilt. Anschließend werden die obigen Parameter durch eine iterative lokale Suche angepasst, um eine bestmögliche Übereinstimmung der thermischen Analyse mittels Faltung (unter Anwendung der Gauss-Funktionen) mit der Referenzlösung zu finden.

Die C++-Implementierung der Methodik sowie Datensätze für experimentelle Untersuchungen sind in unter [10] frei verfügbar.

IV. Experimentelle Untersuchungen

Ein Satz von zehn verschiedenen Bussen [10], zwischen und innerhalb von Chips verlaufend, ist mittels der vorgestellten Methodik erfolgreich in mehreren repräsentativen GSRC-Benchmarks eingebunden worden. Die Busse sind in ihren maximalen Abmessungen beschränkt, verbinden jeweils bis zu fünf Blöcke mit je 64 Leiterbahnen, und sind teilweise überschneidend, d.h. manche Blöcke sind mittels mehreren Bussen zu verbinden. Solch ein komplexes Szenario von Verbindungsstrukturen ist bisher einmalig in der Literatur, daher konnten keine Vergleiche mit anderen Arbeiten erfolgen.

Der relative Vergleich in Tabelle I zu Experimenten ohne gezielter Blockausrichtung zeigt, dass eine geringfügige Verschlechterung der Packungsdichte auftritt. Dies ist zu erwarten, da explizite Blockausrichtungen die Flexibilität der Layoutoptimierung naturgemäß einschränken. Weit wichtiger jedoch ist die Tatsache, dass dank der Blockausrichtungen alle hochparallelen Verbindungsstrukturen effektiv verdrahtbar sind; es ergab sich eine um durchschnittlich 30% verkürzte Globalverdrahtung.

Einen beispielhaften Floorplan nach erfolgreicher Planung einiger hochparalleler Verbindungsstrukturen zeigt Abb. 3.

Weitere Ergebnisse unabhängig von gezielter Blockausrichtung, d.h. bei regulärer Anwendung der Benchmarks, sind in Tabelle II hinterlegt. Die Tabelle verdeutlicht, dass unsere Methodik vergleichbare bzw. bessere Ergebnisse liefert als ein kraftbasiertes Tool [11] bzw. ein SA-basiertes Tool [12]. Beispielsweise erreicht unsere Methodik ähnliche Verdrahtungslängen sowie maximale Temperaturen bei geringerem Flächenverbrauch im Vergleich zu [11]. Des Weiteren ist im Vergleich zu [12] erkennbar, dass unsere Methodik die Verdrahungslängen sowie die maximalen

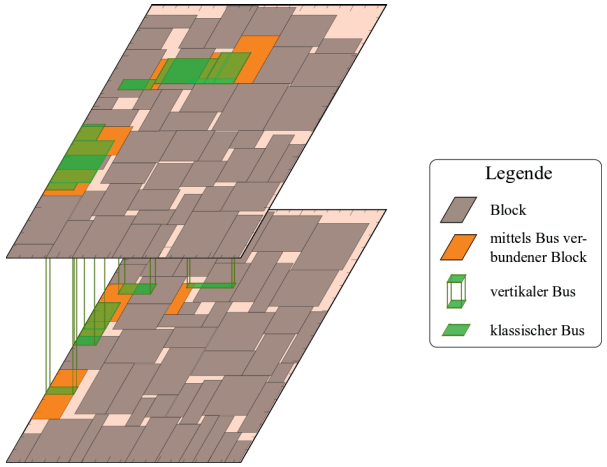

Abbildung 3. Erfolgreich geplante hochparallele Verbindungsstrukturen (Busse) dank entsprechend gezielter Blockausrichtungen für den Benchmark *n100*.

Temperaturen maßgeblich reduzieren kann, unter Nutzung etwas größerer Chipflächen. Damit liegt nahe, dass wesentliche Kompromisse des 3D-Floorplanning, wie Packungsdichte kontra maximaler Temperatur, effektiv adressiert werden.

V. Zusammenfassung

Dieser Beitrag erweitert das 3D-Floorplanning um eine vorausschauende Planung von hochparallelen Verbindungsstrukturen. Solch eine Planung ist eine wichtige, bisher jedoch vor allem in frühen Entwurfsphasen unzureichend berücksichtige Aufgabe. Um diese Aufgabe zu lösen, ist eine gezielte Ausrichtung von (mittels hochparallelen Verbindungsstrukturen gekoppelten) Blöcken vorgeschlagen. Zu Beginn dieses Beitrags erfolgte eine Diskussion, welche Arten von Verbindungsstrukturen anhand eines flexiblen Konzeptes zur Blockausrichtung modellierbar sind. Dieses Konzept wurde anschließend anhand einer neuen 3D-Floorplanning-

Tabelle I
Ergebnisse bei Anwendung der vergrösserten GSRC-Benchmarks. Planung von hochparallelen Verbindungsstrukturen mittels gezielter Blockausrichtung (obere Hälfte) im Vergleich zum regulären Floorplanning (untere Hälfte).

Metrik	2 Chips			3 Chips		
	n100	*n200*	*n300*	*n100*	*n200*	*n300*
Verdrahtung ($cm \times 10^3$)	1,18	1,81	1,97	1,10	1,93	2,07
Fixe Chipmaße (cm^2)	1,14	1,14	1,14	0,73	0,84	0,91
Deadspace (freie Flächen) (%)	29,21	30,39	31,14	26,81	37,20	42,06
Laufzeit (s)	80	359	891	81	360	858
Verdrahtung ($cm \times 10^3$)*	1,83	2,60	2,53	1,34	2,59	2,76
Fixe Chipmaße (cm^2)	1,00	1,08	1,07	0,77	0,82	0,35
Deadspace (freie Flächen) (%)	18,82	27,04	26,39	29,81	36,00	32,19
Laufzeit (s)	59	304	726	59	304	734

*Erhöhter Verdrahtungsbedarf aufgrund von nicht ausgerichteten Blöcken der Verbindungsstrukturen ist berücksichtigt.

J. Knechtel, J. Lienig, S. Osmolovskyi "3D-Floorplanning für hochparallele Verbindungsstrukturen," Tagungsband Dresdner Arbeitstagung Schaltungs- und Systementwurf (DASS 2014), Fraunhofer Verlag, ISBN 978-3-8396-0738-1, pp. 16-20, April 2014.

Tabelle II
VERGLEICHENDE ERGEBNISSE FÜR "KLASSISCHES" 3D-FLOORPLANNING, MIT FOKUS AUF OPTIMIERUNG DER PACKUNGSDICHTE, VERDRAHTUNGSLÄNGE UND MAXIMALER TEMPERATUR.

Metrik	Corblivar, 2 Chips			Corblivar, 2 Chips		Corblivar, 3 Chips			Corblivar, 3 Chips	
	n100	*n200*	*n300*	*ami33*	*xerox*	*n100*	*n200*	*n300*	*ami33*	*xerox*
Verdrahtung ($\mu m \times 10^5$)	3,70	6,57	9,07	2,02	13,89	4,24	7,19	10,28	2,06	16,36
Fixe Chipmaße ($\mu m^2 \times 10^5$)	1,01	0,99	1,61	10,38	143,15	0,75	0,68	1,08	8,98	117,48
Deadspace (%)	11,98	12,01	15,65	44,31	32,41	20,53	14,62	16,27	57,08	45,09
Max. Temp. [9] (°K)	313,81	314,53	315,95	309,14	353,85	355,62	363,94	363,35	333,17	416,61
Laufzeit (s)	108	286	548	50	14	154	380	704	71	22
Metrik	[11], 2 Chips			[12], 2 Chips		[11], 3 Chips			[12], 3 Chips	
	n100	*n200*	*n300*	*ami33*	*xerox*	*n100*	*n200*	*n300*	*ami33*	*xerox*
Verdrahtung ($\mu m \times 10^5$)	3,65	6,18	9,53	1,81	17,14	4,59	7,17	10,61	2,22	21,86
Fixe Chipmaße ($\mu m^2 \times 10^5$)	1,19	1,21	2,15	9,65	125,17	0,97	0,82	1,48	7,54	88,93
Deadspace (%)	25,02	27,87	36,69	40,09	22,70	38,89	28,64	38,65	48,87	27,47
Max. Temp. [9] (°K)	313,31	313,74	314,63	336,36	366,48	348,55	360,60	361,35	384,39	482,16
Laufzeit (s)	439	446	526	193	47	266	497	574	193	48

Methodik umgesetzt. Eine wichtige Erkenntnis ist, dass komplexere Blockausrichtungen, vor allem solche zwischen verschiedenen Chips, eine Synchronisation der Layouterstellung erfordern. Dieser Gegebenheit wird in unserer Methodik anhand einer hierarchischen und orchestrierenden Prozessstruktur Rechnung getragen. Die Implementierung der 3D-Floorplanning-Methodik erfolgt unter Anwendung und Erweiterung der bekannten Simulated-Annealing-Heuristik, als auch einer effektiven und akkuraten thermischen Analyse, genannt Power Blurring. Experimentelle Untersuchungen unter Anwendung der GSRC-Benchmarks demonstrieren das Potential unserer Methodik sowohl für die gezielte Planung von hochparallelen Verbindungsstrukturen als auch für Aspekte des "klassischen" 3D-Floorplanning. Letztere sind vor allem die Optimierung der Packungsdichte und das thermische Management.

ANMERKUNGEN

Gefördert durch die DFG (Projekt 1401/1). Inhalte dieses Beitrages wurden unter Zusammenarbeit mit Prof. Evangeline Young (Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong) erstellt und sind in [13] einsehbar.

LITERATUR

[1] G. H. Loh *et al.*, "Processor design in 3D die-stacking technologies," *Micro*, vol. 27, pp. 31–48, 2007.
[2] D. H. Kim *et al.*, "3D-MAPS: 3D massively parallel processor with stacked memory," in *Proc. Int. Solid-State Circ. Conf.*, 2012, pp. 188–190.
[3] H. Xiang *et al.*, "Bus-driven floorplanning," in *Proc. Int. Conf. Comput.-Aided Des.*, 2003, pp. 66–73.
[4] J. H. Y. Law and E. F. Y. Young, "Multi-bend bus driven floorplanning," *Integration*, vol. 41, no. 2, pp. 306–316, 2008.
[5] J. Knechtel *et al.*, "Assembling 2-D blocks into 3-D chips," *Trans. Comput.-Aided Des. Integr. Circuits Sys.*, vol. 31, no. 2, pp. 228–241, 2012.
[6] D. H. Kim *et al.*, "Block-level 3D IC design with through-silicon-via planning," in *Proc. Asia South Pacific Des. Autom. Conf.*, 2012, pp. 335–340.
[7] X. Hong *et al.*, "Corner block list: an effective and efficient topological representation of non-slicing floorplan," in *Proc. Int. Conf. Comput.-Aided Des.*, 2000, pp. 8–12.
[8] J.-H. Park *et al.*, "Fast thermal analysis of vertically integrated circuits (3-D ICs) using power blurring method," in *Proc. ASME InterPACK*, 2009, pp. 701–707.
[9] A. Coskun *et al.* (2011) Hotspot 3D extension. [Online]: http://lava.cs.virginia.edu/HotSpot/links.htm
[10] J. Knechtel. (2014) Corblivar floorplanning suite. [Online]: http://www.ifte.de/english/research/3d-design/index.html
[11] P. Zhou *et al.*, "3D-STAF: scalable temperature and leakage aware floorplanning for three-dimensional integrated circuits," in *Proc. Int. Conf. Comput.-Aided Des.*, 2007, pp. 590–597.
[12] Y. Chen. (2010) 3DFP – thermal-aware floorplanner for three-dimensional ICs. (Eine dazugehörige Publikation ist in [14] einsehbar.). [Online]: http://www.cse.psu.edu/~yxc236/3dfp/index.html
[13] J. Knechtel *et al.*, "Structural planning of 3D-IC interconnects by block alignment," in *Proc. Asia South Pacific Des. Autom. Conf.*, 2014, pp. 53–60.
[14] W.-L. Hung *et al.*, "Interconnect and thermal-aware floorplanning for 3D microprocessors," in *Proc. Int. Symp. Quality Elec. Des.*, 2006, pp. 98–104.

J. Knechtel, J. Lienig, S. Osmolovskyi "3D-Floorplanning für hochparallele Verbindungsstrukturen," Tagungsband Dresdner Arbeitstagung Schaltungs- und Systementwurf (DASS 2014), Fraunhofer Verlag, ISBN 978-3-8396-0738-1, pp. 16-20, April 2014.