

Inter-Processor communication patterns in weather forecasting models

Tomas Wilhelmsson

Swedish Meteorological and Hydrological Institute

Sixth Annual Workshop on

Linux Clusters for Super Computing

2005-11-18

SMHI Numerical Weather Prediction

Analysis

Obtain best estimate of current weather situation from

2. Background, (the last forecast 6 to 12 hours ago)

3. Observations (ground, aircraft, ship, radiosondes, satellites)

Variational assimilation in 3D or 4D

Most computationally expensive part

Forecast

Step forward in time (48 hours, 10 days, ...)

Ensemble forecast

Estimate uncertanty by runing many (50-100) forcasts from perturbed analysis

A 10-day ensemble forecast for Linköping

• Blue line is unperturbed high resolution forecast

- Dotted red is unperturbed reduced resolution forecast
- Bars indicate center 50% of 100 perturbed forecasts at reduced resolution

EPS Meteogram Linkoping 57.9° N 15.6° E 33N Deterministic Forecasts and EPS Distribution 17 October 2005 12 UTC Total Cloud Cover (okia)

7.65



----- TL255 CTRL ------ TL511 CIRS



48-hour 22 km resolution forecast on a limited domain

-Boundaries from global IFS forecast at 40km

Also 11 km HIRLAM forecast on a smaller domain

• 40 minutes elapsed on 32 processor of a Linux cluster

– Dual Intel Xeon 3.2 GHz

- Infiniband

– More info in Torgny Faxén's talk tomorrow!



SMHI Codes: IFS, ALADIN, HIRLAM, ALADIN

- IFS Integrated Forecast System (ECMWF)
 - Global, Spectral, 2D decomposition, 4D-VAR
- ALADIN Aire Limitée Adaptation Dynamique développement InterNaltional
 - Shares codebase with ARPEGE, the Météo-France version of IFS
 - Limited area, Spectral, 2D decomposition, 3D-VAR
 - Future: AROME at 2-3 km scale
- HIRLAM High Resolution Limited Area Model
 - Limited area, Finite difference, 2D decomposition
- HIRVDA HIRIam Variational Data Assimilation
 - Limited area, Spectral, 1D decomposition, 3D-VAR, (and soon 4D-VAR ©)



Longer time steps made possible by

•Semi-implicit time integration

- Advance fast linear modes implitly and slower non-linear modes explicitly
- A Helmholz equation has to be solved
 - In HIRLAM by direct FFT + tri-diagonal method
 - Spectral models do it easily in Fourier space
- Implications for domain decomposition!

•Semi-Lagrangian advection

- Wide halo zones



SMHI How should we partition the grid?

- Example: HIRLAM C22 grid (nx = 306, ny = 306, nlev = 40)
 - Many complex interactions in vertical (the "physics").
 - Decomposing the vertical would mean frequent interprocessor communication.
 - Helmholtz solver
 - FFT part prefers nondecomposed longitudes
 - Tridiagonal solver partpreferes nondecomposed latitudes
- Similar for spectal models (IFS, ALADIN & HIRVDA)
 - Transforming from physical space to spectral space means
 - FFTs in both longitiudes and latitudes
 - And physics in vertical

SMHI Grid partitioning in HIRLAM (Jan Boerhout, NEC)





SMHI

Transforms and transposes in IFS / ALADIN



Signatur



Spectral methods in limited area models HIRVDA / ALADIN



SMHI Transposes in HIRVDA (spectral HIRLAM) 1D decomposition







SMHI Transposes with 2D partitioning



SMHI Load balancing in spectral space

•Isotropic representation in spectral space requrires an ellipic truncation

• By accepting an unbalanced y-direction FFT, spectral space can be load balanced





1D decomposition

$$2n(n-1)$$

2D decomposition

$$\frac{1}{6}6n(\sqrt{n}-1)$$

Signatur

SMHI Timings on old cluster (Scali)



SMHI Timings on new cluster (Infiniband)







SMHI Minimum time on old cluster



SMHI FFT / Transpose timeline 2D decomposition

		2,37 s		2.37 <u>5 s</u>		2.38 s	2.385 s		2.39 s	
ocess 0	248 223	516	223	520	223 526	API Waitali	530	223 TRA	CE_OFF	MPI
ocess 1	248	516	223	520	526	MPI Waitali	xft inv		CE OFF	ym_
ocess 2	240	516	X23	520	526				ACE OFF	- V#-
ocess 3	240	510		520	526		xm_inv		RACE OFF	tran
ocess 4	240			520	520	A DATE OF STREET		500		trans
DCess 5	240	INI Hall E4R		520	520	A CLASSES AND A CLASSES	JAN 1530		TRACE OFF	trans
Juess 0	240	Maltell 516		520	526					trans
ncess 7	240			520	222 526		530	223 TP	ACE OFF	trap
cess 0	240 223	516		520	1222 526 SI					xfft
cess 1	248	516		520	528				RACE OFF	xfft
cess 11	248	516		520	526		530		TRACE OFF	trans
cess 17	248 523	516		520	526		wifft inv		TRACE OFF	t_tes
cess 12	248	-1516		520	526		A A A A A A A A A A A A A A A A A A A		TRACE OFF	
cess 1/	248	516		AND I AND	526		wift inv		TRACE OFF	
cess 15	248 223	516		520	526		AND AND AND	1112000	TRACE OFE	
cess 1f	248 223	516		520 223	3 - 223 - 526 MPI		A A A A A A A A A A A A A A A A A A A	223	TRACE OFF	
cess 17	248	516		520	223 526			223	ACE OFF	
cess 1F	248	516			526			223	RACE OFF	
ess 19	248	516	A STREET STREET	520	526		530		RACE OFF	
cess 20	248	516		520	526		530		TRACE OFF	
cess 21	248	516		520	526		530	1 - 70 / A	TRACE OFF	
cess 22	243	516		520	526		xfft inv		TRACE OFF	
cess 2?	248 223	516		520	223 526			223	TRACE OFF	
cess 24	248 223/	516		520	222 526		xfft inv	223	TRACE OFF	
cess 25	243 223	516		520	526			223	RACE OFF	
cess 2f	3 248 223	516		223	526				TRACE OFF	
cess 27	243	516		520	526		530		TRACE OFF	
cess 28	s 🗾 📶 🚟 🏹	516	HARA AND NO	520	526		xfft_inv		TRACE OFF	
cess 29	248	516		520	526		xfft_inv		TRACE OFF	
cess 30	243	51 6 /		520	526		530		TRACE OFF	
cess 31	249 223	516 4	Contraction and the second second	520 223	223 526		nv inv	11/- 223	TRACE OFF	
cess 32	248 223	516	New York A. C.	520	223 223 526		xfft_inv	223	TRACE_OFF	
cess 33	i 248 - 283	516		520	223 526		200 (Contraction of the contrac	MPL Waitai	RACE_OFF	
ess 34		See 200 516		520	< Sec. 1 / 526 / 526 / / 200 / / 200 / / 200 / 2	All the second second		P Note	TRACE_OFF	
cess 35	i <mark>248 - 200</mark>	516		520	526 5 26	Stor Vieland	530		TRACE OFF	
cess 36	; 🗾 🗖 🚟	516	14 An 23 St. St. St.	520	526 S	A Barren States	A A A A A A A A A A A A A A A A A A A	-223 /////	TRACE OFF	
ess 37	248	516		520	526		xfft_inv	17792255	TRACE OFF	
ess 38		516	Martin Carlos States and	520	526 /	March States	xfft_inv		TRACE OFF	
cess 39	243 223	516	Merel Carlos Andrews	520	223 526 ///	ACC DESCRIPTION AND	inv	223	TRACE OFF	
ess 40	1 248 223	516	January of Science and and the		526	Allen Alle College and Anna	xfft_inv	223	TRACE OFF	
ess 41	248	516	HILL STREET	520	526	MALLAND STATE	530	MPI Waitah	TRACE OFF	
ess 42	240	516	and the second of the second second	520	526	ALL A LEWISCH AND			TRACE OFF	
cess 43	1 1/413	516	Marking Shiekan Salasan	520	526	A CARA CARA CARA CARA CARA CARA CARA CA	530		TRACE OFF	
ess 44	499	516	1/ Indiana Anna		526	The shall be	A CARLON AND A		TRACE OFF	
cess 45	2-18	516			526	The second state	xff_inv		TRACE OFF	
ess 46	1 21919	516	San State Association		526	MIT ALL MARKS	xfft_inv		TRACE OFF	
ess 47	243 223	516			526		inv inv	223	TRACE OFF	
ess 48	1 246 223	516	And Market Constant	520	526		xfft_inv	223	TRACE OFF	
ess 49		516		520	526			WPI Walter		
ess 50		516		620	526				RACE OFF	
ess 51					526	MIL MARKA	inv_		TIRACE OFF	
10ee 57		516								
	243	516		620	526		THE REAL PROPERTY OF THE PARTY		TRACE OFF	
cess 53	243 243	516 516 516		520 520	526 526		530		TRACE OFF	
ess 53 ess 54	243	516 516 516		520 HF	526 526 526		530		TRACE OFF TRACE OFF	
cess 53 cess 54 cess 55	243 240 248 248 223	516 516 516 516		520 520 520 11	526 526 526 526 526 526 526 526	MP: Waita	530 Signature Signat	223	TRACE OFF TRACE OFF TRACE OFF	
cess 51 cess 54 cess 55 cess 55	2 243 223 243 223 248 223 248 223 248 223	516 516 516 516 516 516	MP Waital	520 Hiff	526 526 526 526 526 526 526 526 526 526	MP Vai MP Vai MP Vai MP Vai MP Vai MP Vai MP Vai MP Vai	530 'sfit'in sfit'in sitt'in sitt'in	223	TRACE OFF TRACE OFF TRACE OFF TRACE OFF	

SMHI FFT / Transpose timeline **1D decomposition**

			1.6	85 s		1.69 s	1.69	5 s	1.7 s	
ocess 0	TRACE OFF	2.8512	xfft_dir		223	518 520	223	524	TRACE OFF	MPI
cess 1	TRACE_OFF	512	514			518 520	MPL Waiteli	za xfft	_inv TRACE_OFF	yfft_dir
cess 2	TRACE OFF	243 512	xfft_dir			518 520		<u>524</u>	TRACE OFF	yfft_inv
cess 3	TRACE_OFF	<mark>- 512</mark>	514		AND DISIDA	518 520 -		nv nv	TRACE_OFF	VI_API
cess 4	TRACE OFF	243 512	xfft_dir			518 520		V 📈 🌅 <mark>5</mark> 2	4 TRACE OFF	trans dir
cess 5	TRACE_OFF	512	xfft_dir			18 520 🦰		· · · // _//	TRACE OFF	
cess 6	TRACE OFF	243 512	xfft_dir			518 520			TRACE OFF	
cess 7	TRACE_OFF	512	514			18 520		linv <u>(</u> inv	TRACE_OFF	trans in
cess 8	TRACE OFF	512	514			518 520		🖉 🖕 🕺 🕺 🖊	inv TRACE OFF	t_test
cess 9	TRACE_OFF	243 512	xfft_dir	-		518 520		524	TRACE OFF	
cess 10	TRACE OFF	243 512	xfft_dir			518 520 🥻	D	524	TRACE OFF	
cess 11	TRACE_OFF	248 512	xfft_dir	100		518 520 🧹	Pro 1997	52	4 TRACE_OFF	
cess 12	TRACE OFF	243 512	xfft_dir	110		518 520	1	524	TRACE OFF	
cess 13	TRACE OFF	512	514			518 520 🦉		l 🔬 🔊	TRACE OFF	
cess 14	TRACE OFF	248 512	xfft_dir	1 Alexandre		518 520		5	24 TRACE OFF	
cess 15	TRACE OFF	512	514			520			TRACE OFF	
cess 16	TRACE OFF	243 512	xfft_dir	1000 m		518 520		24	TRACE OFF	
cess 17	TRACE OFF	512	514			18 520	te a secondaria de la companya de la		TRACE OFF	
cess 18	TRACE OFF	243 512	xfft_dir	- Min		518 520		524	TRACE OFF	
cess 19	TRACE_OFF	248 512	xfft_dir	1110		518 520 🧖		524	TRACE OFF	
cess 20	TRACE OFF	512	514			518 520		tfft_inv	TRACE_OFF	
cess 21	TRACE OFF	248 512	xfft_dir	V ile		518 520			TRACE OFF	
cess 22	TRACE OFF	248 512	xfft_dir	Marke		518 520	21	524	TRACE OFF	
cess 23	TRACE_OFF	512	514			8 520		<u>av</u>	TRACE OFF	
cess 24	TRACE OFF	243 512	xfft_dir	1111		518 520		524	TRACE OFF	
cess 25	TRACE OFF	512	514			3 520			TRACE OFF	
cess 26	TRACE OFF	248 512	xfft_dir			518 520		524	TRACE OFF	
cess 27	TRACE OFF	243 512	xfft_dir	- Marin		518 520		524	TRACE OFF	
cess 28	TRACE OFF	248 512	xfft_dir	- Cherry		518 520		524	TRACE OFF	
cess 29	TRACE OFF	512	514			520			TRACE OFF	
cess 30	TRACE OFF	<mark>/=</mark> 512	514	V		518 520		xfft_i	NV TRACE OFF	
cess 31	TRACE_OFF	243 512	xfft_dir	1		518 520 7		524	TRACE OFF	
cess 32	TRACE OFF	512	xfft_dir	ll a s		518 520		524	TRACE OFF	
cess 33	TRACE_OFF	248 512	xfft_dir			518 520	1		TRACE OFF	
cess 34	TRACE OFF	512	514			518 520		xfft	_inv TRACE_OFF	
cess 35	TRACE_OFF	243 512	xfft_dir	i de la composición de la comp		518 ¹ 520		24	TRACE OFF	
cess 36	TRACE OFF	248 512	xfft_dir			518 520		524	TRACE OFF	
cess 37	TRACE_OFF	512	514	AN ST I		518 520		xfft	inv TRACE_OFF	
cess 38	TRACE OFF	248 512	xfft_dir	lini i		518 520		<u>24</u>	TRACE OFF	
cess 39	TRACE_OFF	<mark>512</mark>	514			518 520 📕		<u> </u>	TRACE OFF	
cess 40	TRACE OFF	248 512	514	V lini		518 520		52	4 TRACE_OFF	
cess 41	TRACE_OFF	240 512	xfft_dir	Nil.		518 520		52	4 TRACE_OFF	
cess 42	TRACE OFF	512	514	1 killing		518 520		<u>nv</u>	TRACE OFF	
cess 43	TRACE_OFF	243 512	xfft_dir			518 520		624	TRACE_OFF	
cess 44	TRACE OFF	243 512	xfft_dir			518 520		524	TRACE_OFF	
cess 45	TRACE_OFF	512	514	1 in		18 520 119		<u>t inv</u>	TRACE_OFF	
cess 46	TRACE OFF	243 512	xfft_dir			518 520///		524	TRACE_OFF	
cess 47	TRACE_OFF	512	514	Ale -		518 520		inv 💦	TRACE OFF	
cess 48	TRACE OFF	248 512	xfft_dir			518 520		524	TRACE OFF	
cess 49	TRACE_OFF	243 512	xfft_dir	11 12	1 Martin State State	518 520		524	TRACE OFF	
cess 50	TRACE OFF	248 512	514		and the second	518 520	and the second states	524	TRACE OFF	
cess 51	TRACE_OFF	243 512	514		a far a sea a far a f	518 520 🦛	المالية والمعالية المتحر المحرج ا	xfft_inv	TRACE OFF	
cess 52	TRACE OFF	248512	514		111 J. C. C. S.	518 520 🗖	A MARCONS	xfft_inv	TRACE OFF	
cess 53	TRACE OFF	248 512	514		A Carlos Carlos	518 520	11 14 23 243	524	TRACE OFF	
cess 54	TRACE OFF	248 512	514		1	518 520		524	TRACE OFF	
cess 55	TRACE OFF	248 512	514	1. and		518 520		524	TRACE OFF	
cess 56	TRACE OFF	243 512	514	histol		518 520		xfft_inv	TRACE OFF	
cess 57	TRACE OFF	248 512	514	1 Maria	MPI Waltell	518 520		524	TRACE OFF	
cess 58	TRACE OFF	248 512	514		MPI Waitall	518 520	MPI Waitall	xfft inv	TRACE OFF	59/64 bars

Printed by Intel(R) Trace Inalyzer 4

SMHI Semi-Lagrangian Advection



• Full cubic interpolation in 3D is 32 points (4x4x4)



Example: The HIRLAM C22 area (306x306 grid at 22 km resolution)

- Max wind speed in jet stream 120 m/s
- Time step 600 s
- => Distance 72 km = 3.3 grid points)
- Add stencil width (2) => nhalo= 6
- With 64 processors partitioned in 8x8:
 - -38x38 core points per processor
 - 50x50 including halo
- Halo area is 73% of core!
- But full halo is not needed everywhere!





IFS & ALADIN – Semi-Lagrangian advection Requesting halo points 'on-demand'





- 1. Exchange full halo for wind components (u,v & w)
- 2. Calculate departure points
- 3. Determine halo-points needed for interpolation
- 4. Send list of halo points to surrounding PE's
- 5. Surrounding PE's send points requested

Effects on various optimizations on IFS performance



Moving from Fujitsu VPP (vector machine) to IBM SP (cluster).

Figure from Debora Salmond (ECMWF).

Signatu



• Meteorology and climate sciences provide plenty of fun problems for somebody interested in computational methods and parallelization. Also:

- Load balancing observations in data assimilation
- Overlapping I/O with computation
- HIRLAM & HIRVDA will get 'on-demand' slswap soon